MEGI

MESTRADO
Estatística e Gestão de Informação
Master in Statistics and Information Management

DETERMINAÇÃO DA PREVALÊNCIA DE FIBRILHAÇÃO AURICULAR NA POPULAÇÃO PORTUGUESA AO NÍVEL DA NUTS III USANDO MODELOS HIERÁRQUICOS BAYESIANOS

Nilza Karina Cabral Gonçalves

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Estatística e Gestão de Informação, Especialização em Marketing Research e CRM

UNIVERSIDADE NOVA ISEGI Estatística e Gestão de Informação

Instituto Superior de Estatística e Gestão de Informação
School of Statistics and Information Management
Universidade Nova de Lisboa

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

DETERMINAÇÃO DA PREVALÊNCIA DE FIBRILHAÇÃO AURICULAR NA POPULAÇÃO PORTUGUESA AO NÍVEL DA NUTS III USANDO MODELOS HIERÁRQUICOS BAYESIANOS

por

Nilza Karina Cabral Gonçalves

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Estatística e Gestão de Informação, Especialização em Marketing Research e CRM

Orientador: Professor Doutor Jorge M. Mendes

Fevereiro, 2014

AGRADECIMENTOS

Learn from yesterday, live for today, hope for tomorrow. The important is not to stop questioning.

(Albert Einstein)

Este trabalho resulta de um caminho que fiz acompanhada por muita gente a quem agradeço, mas não podia deixar de referir o meu orientador, Professor Jorge M. Mendes, pela paciência no ensinamento, pela disponibilidade demonstrada, pelo incentivo gerado para continuar a questionar, aos meus pais, Cláudio e Gilda, pelo esforço de vida que tiveram e pela educação que me deram, fizeram com que eu chegasse até aqui e não deixe de continuar a questionar. Ao Nelson, porque mesmo sem saber foi sempre eficaz nos seus comentários.

À Catarina, pela disponibilidade para as correcções e comentários. Ao Instituto Português do Ritmo Cardíaco, cuja cedência dos dados permitiu a conceptualização deste trabalho.

E, por último como se fosse o primeiro, ao Vitor o meu agradecimento eterno, porque sempre acreditou e mesmo quando a motivação era baixa ele estava lá a incentivar, a apoiar e a cuidar para este projecto chegasse a bom porto.

RESUMO

A Fibrilhação Auricular é uma alteração do ritmo cardíaco designada por arritmia. Esta patologia é considerada a forma de arritmia mais frequentemente observada na prática clínica e que constitui uma importante causa de morbilidade pelo risco inerente de desenvolvimento de AVC. Em 2010 foi realizado um estudo epidemiológico na população Portuguesa com o objectivo determinar a prevalência de Fibrilhação Auricular na população portuguesa com idade igual ou superior a 40 anos, sob o acrónimo de FAMA. Os dados publicados indicaram uma estimativa de prevalência de 2,5%, com um aumento da prevalência em função da classe etária. A nível regional não foram observadas diferenças na taxa de prevalência. Estudos de mapeamento de doenças mostraram que a determinação de taxas de prevalência por região, quando o número de casos observados é relativamente baixo, apresentam sobredispersão e, consequentemente, uma falta de precisão nas estimativas obtidas através um método frequencista clássico. A utilização de modelos Bayesianos hierárquicos no mapeamento de doenças tem apresentado vantagem na estimação de valores de risco da doença comparativamente à abordagem clássica. Assim, é objectivo deste trabalho determinar a prevalência de Fibrilhação Auricular na população Portuguesa por região, ao nível da NUTS III, usando modelos hierárquicos Bayesianos.

Os dados utilizados neste estudo são os dados referentes ao estudo FAMA, pós-estratificados para correcção dos ponderadores. O modelo Bayesiano proposto por Besag, York e Mollié (1991) foi usado para modelar os dados, covariando para a idade e índice de massa corporal. A revisão do desenho do estudo e o recálculo dos ponderadores foi realizado com recurso ao *software* R, *survey*, e a análise Bayesiana efectuada através do *WinBugs*.

Os resultados deste estudo demonstram que o uso de modelos Bayesianos são uma melhor opção para a estimação de valores risco relativo e prevalência da doença. Contudo, a utilização de covariáveis não resultou numa melhoria considerável ao contrário do que seria esperado. Conclui-se que a Fibrilhação Auricular apresenta variações regionais significativas, a nível de NUTS III, que não devem ser desvalorizadas na determinação de políticas de saúde pública para controle da doença.

PALAVRAS CHAVE

Fibrilhação Auricular; Prevalência; Modelos Hierárquicos Bayesianos; Fully Bayesian; Modelos CAR

ABSTRACT

Atrial fibrillation is an abnormal heart rhythm called arrhythmia. This condition is considered the most common form of arrhythmia observed in clinical practice and its a major cause of morbidity due the inherent risk of stroke. In 2010 an epidemiological study was conducted in the Portuguese population in order to determine the prevalence of atrial fibrillation in the Portuguese population aged over 40 years, under the acronym FAME. Published data indicated an estimated prevalence of 2.5%, with an increased prevalence depending on the age group. No regionally differences were observed in prevalence rate. Disease mapping studies have shown that the prevalence estimation by region, when the number of observed cases is relatively small, has an overdispersion and thus a lack of accuracy in the estimates obtained by the classical method. The use of Bayesian hierarchical models in disease mapping has shown benefit in the estimation of values at risk of the disease compared to the classical approach. Thus, the objective of this work was to determine the prevalence of atrial fibrillation in the Portuguese population by region, NUTS III level, using Bayesian hierarchical models.

The data used in this study is the data concerning the FAME study, post-stratified for weights correction. The hierarchical Bayesian model proposed by Besag, York and Mollie (1991) was used to model the data, adding age and body mass index as covariates to the model. The review of the study design and the recalculation of the weights was performed using the R software, package 'survey', and Bayesian analysis through WinBugs.

The results of this study demonstrate that the use of Bayesian models are a better choice for the estimation of relative risk values and prevalence of the disease. However, the use of covariates did not result in a considerable improvement for the estimates precision, contrary to what

one would expect. We conclude that the prevalence of Atrial Fibrillation has significant regional variations, NUTS III level, which should not be undervalued in determining public health policy to control the disease.

KEYWORDS

Atrial Fibrillation; Prevalence; Hierarchical Bayesian Models; Fully Bayesian; CAR models

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Siglas e Abreviaturas

APAPE Associação Portuguesa de Arritmologia, Pacing e Electrofisiologia

AVC Acidente Vascular Cerebral

BIC Bayesian Information Criterion

BUGS Bayesian inference Using Gibbs Sampling

CAR Conditional AutoRegression

DIC Deviance Information Criteria

DM Diabetes Mellitus

DPOC Doença Pulmonar Obstrutiva Crónica

EB Empirical Bayes

ECG Electrocardiograma

EQM Erro Quadrático Médio

EUA Estados Unidos da América

FA Fibrilhação Auricular

FAMA Estudo de Avaliação da Prevalência de Fibrilhação Auricular em Portugal

fdp Função Densidade de Probabilidade

FB Fully Bayes

HTA Hipertensão Arterial

iid independentes e identicamente distribuídos

IMC Índice de Massa Corporal

INE Instituto Nacional de Estatística

IPRC Instituto Português do Ritmo Cardíaco

LVT Lisboa e Vale do Tejo

MCMC Monte Carlo via Cadeias de Markov

MMC Método de Monte Carlo

NUT Nomenclaturas de Unidades Territoriais

OMS Organização Mundial de Saúde

RAA Região Autónoma dos Açores

RAM Região Autónoma da Madeira

RR Risco Relativo

SE Standard Error

ACRO Standardized Disease Ratio

SMR Standardized Mortality Ratio

SPC Sociedade Portuguesa de Cardiologia

TMP Taxa de Mortalidade Padronizada

Capítulo 1

Introdução

1.1 A Fibrilhação Auricular

A Fibrilhação Auricular (FA) é uma patologia do foro cardíaco e que se caracteriza por ser uma arritmia, ou seja, uma alteração do ritmo cardíaco. Esta patologia é considerada a forma de arritmia mais frequentemente observada na prática clínica.

O coração é um órgão que está permanentemente em acção de contracção e relaxamento para que o sangue seja bombeado para todo o organismo. As quatro cavidades que compõem este órgão (aurícula direita, ventrículo direito, aurícula esquerda e ventrículo esquerdo) permitem que o fluxo sanguíneo ocorra num sistema equilibrado, no qual o sangue é bombeado para dentro da aurícula direita através da veia cava superior, passando para o ventrículo direito e saindo pela artéria pulmonar para os pulmões de modo a libertar dióxido de carbono e proceder a nova oxigenação, enquanto no lado contrário o sangue entra pela veia pulmonar para a aurícula esquerda, passa para o ventrículo esquerdo e sai pela aorta, levando sangue oxigenado a todas as partes do corpo. O ritmo de entrada e saída de sangue das cavidades do

coração é controlado pelo nódulo sino-auricular, um conjunto de células que enviam impulsos eléctricos ao músculo cardíaco para este seja estimulado a contrair e relaxar de uma forma coordenada, permitindo que o fluxo sanguíneo ocorra sem problemas. Num coração saudável, os impulsos eléctricos ocorrem num ritmo regular com cerca de 70 pulsações por minuto, mas num doente com FA os impulsos eléctricos são irregulares e, consequentemente, a contracção do músculo cardíaco torna-se descoordenada. Nos indivíduos com FA, o ritmo cardíaco pode ser muito elevado, muito lento ou irregular, e devido a esta situação as cavidades cardíacas não esvaziam por completo o sangue bombeado, nomeadamente das aurículas, e como tal o sangue que permanece dentro destas cavidades coagula. Os coágulos formados no interior destas cavidades, se libertados para o sistema circulatório, podem deslocar-se para outros órgãos através do sistema circulatório, nomeadamente para as artérias do cérebro, potenciando casos de embolia e Acidente Vascular Cerebral (AVC). Desta forma, a FA constitui uma importante causa de morbilidade pelo risco inerente de desenvolvimento de AVC.

Um dos primeiros registos históricos de que se tem conhecimento sobre FA foi feito pelo Imperador Huang Ti Nei Ching Su Wên, que terá governado a China algures entre 2598 e 1696 a.C. (Ferreira, 2008). Com a invenção do estetoscópio, o registo de situações de arritmia tornou-se mais frequente, permitindo o estudo desta patologia relativamente à sua etiologia. Os avanços na electrofisiologia cardíaca (ciência que estuda e explica a actividade eléctrica do coração) permitiram um ainda maior desenvolvimento do conhecimento sobre FA, nomeadamente sobre os mecanismos de condução eléctrica na FA (Ferreira, 2008).

Actualmente sabe-se que existem factores de risco para vir a desenvolver FA durante a vida.

Entre esses factores encontram-se os factores demográficos e antropométricos como a idade, o sexo, raça e estatura, e factores de risco clínico como a presença de hipertensão arterial (HTA), diabetes *mellitus* (DM), doença tiróidea, síndrome metabólico, apneia do sono e disfunção sistólica (Kannel & Benjamin, 2008).

Relativamente ao factores demográficos sabe-se que: a) a FA aumenta com a idade e que a partir dos 50 anos o número de casos com diagnóstico de FA duplica por cada década de vida (Kannel & Benjamin, 2008; Whilhelmsen, Rosengren & Lappas, 2001), b) os homens apresentam 1,5 vezes maior risco de ter FA comparativamente às mulheres, e c) o risco de desenvolver FA nos indivíduos de raça negra é menos de metade comparativamente ao da população caucasiana (Ferreira, 2008). Estudos mais recentes indicam que a diferença na prevalência de FA entre homens e mulheres não se comprova (Ferreira, 2008). Já no que se refere ao factor idade, a explicação do aumento exponencial da prevalência de FA com a idade está directamente relacionada com o envelhecimento e o comprometimento da função cardíaca. Quanto à estatura, vários estudos (Ferreira, 2008) demonstraram que a prevalência de FA é significativamente maior nos indivíduos de estatura elevada por causa da pretensa associação entre a altura e o tamanho da aurícula esquerda (Ferreira, 2008). Todos os factores risco clínico potenciam o desenvolvimento de FA.

O diagnóstico de FA é feito com base nos dados da história clínica do indivíduo e na realização de um exame electrocardiográfico, geralmente electrocardiograma (ECG) que permite classificar a FA em paroxística, persistente ou permanente. A FA paroxística e a FA persistente caracterizam-se por episódios que se resolvem espontaneamente sem a intervenção de

qualquer espécie; a diferença entre as duas refere-se ao facto da FA persistente apresentar uma maior duração comparativamente a um caso de FA paroxística. Na FA permanente só há retorno ao ritmo cardíaco normal com intervenção farmacológica.

Com o envelhecimento da população, o aumento do risco cardiovascular, da prevalência de insuficiência cardíaca, obesidade, DM e síndrome metabólico (Kannel & Benjamin, 2008) projecta-se o cenário ideal para uma epidemia de FA.

1.2 Incidência e prevalência de Fibrilhação Auricular

Um artigo publicado em 2010 por Bonhorst e colegas fez uma revisão de literatura dos dados de incidência (número de novos casos da patologia em estudo identificados numa dada população num determinado período de tempo) e prevalência (número total de casos observados numa dada população e num determinado período de tempo) da FA publicados nos Estados Unidos da América (EUA), na Europa e em Portugal. Nos EUA as taxas de incidência variam entre 2 e 3 por 1.000 pessoas/ano e 22,4 por 1.000 pessoas/ano. Na Europa, os valores de incidência variam entre os 0,9 por 1.000 pessoas/ano e 9,9 por 1.000 pessoas/ano, com diferenças substanciais nos valores observados entre a população do sexo masculino e feminino.

A revisão dos valores de prevalência de FA também mostrou variações consideráveis. Nos EUA a prevalência de FA variou entre os 0,95%, em indivíduos com mais de 20 anos, e os 6,2% para uma população com idade $\geq$ 65 anos. Para a Europa os valores de prevalência apresentados variam entre <1%, num estudo realizado na Escócia numa população de indivíduos com idade compreendida entre os 45 e 64 anos, e os 8,5% referente a um estudo

desenvolvido em Espanha junto de uma amostra de 7.108 indivíduos pertencentes a uma população de homens e mulheres com >60 anos.

A conclusão desta revisão refere que a incidência e prevalência desta arritmia tem variações geográficas significativas mas algumas limitações metodológicas condicionam a interpretação e comparação destes resultados. Por um lado, o uso de diferentes metodologias para a recolha de dados impede uma comparação directa dos resultados. De facto, todos os estudos apresentados reflectem metodologias distintas na recolha de informação, uns obtiveram informação através da revisão de processos clínicos e consequentemente as estimativas apresentadas são referentes a uma população hospitalar, enquanto outros identificaram diagnósticos de FA através da realização de ECG e aplicação de questionários a uma amostra populacional. Por outro lado temos para diferentes estudos diferentes definições da população em estudo. Quando comparados os limites etários, observa-se que a idade mínima varia entre os 45 e os 55 anos. Numa patologia onde a idade é um factor preditor da doença, esta variação pode comprometer a interpretação das estimativas de incidência e prevalência da doença.

No entanto, e apesar desta limitação, é consensual entre os vários estudos apresentados, seja de incidência ou prevalência, que as estimativas de FA podem duplicar a partir dos 50 anos por cada década de vida (Kannel & Benjamin, 1998). Segundo Kannel e Benjamim (1998) a prevalência de FA varia entre 0,5% em indivíduos com 50-59 anos até cerca de 9% em indivíduos com 80-89 anos. Entre 1980 e 2000 foi realizado um estudo de coorte numa população do Minnesota (Miyasaka, Barnes, Gersh, et al., 2006) com o objectivo de avaliar a evolução da incidência e prevalência de FA nesta população durante este período. De acordo com este

estudo, em 1980 a prevalência de FA ajustada ao sexo e idade era de 3,04% e em 2000 de 3,68% por 1.000/pessoas ano. Através da modelação dos dados observou-se que a prevalência estimada de FA, em 1990, era de 1,9% em indivíduos com idade compreendida entre os 45 e 64 anos, 6,2% entre os 65 e 74 anos e 13,5% em indivíduos com 75 ou mais anos. Em 1995 Feinberg e colegas fizeram uma revisão dos dados epidemiológicos sobre FA com o objectivo de estimar o número de casos de FA e a sua distribuição por sexo e idade. Foram analisados dados provenientes de quatro grandes estudos, tendo sido estimado uma prevalência inferior a 1% para indivíduos com idade inferior a 59 anos, entre 1% e 3% em indivíduos com idade entre os 60 e 69 anos, entre os 5% e 7% em indivíduos entre os 70 e 79 anos, e igual ou acima dos 10% em indivíduos com 80 ou mais anos. Mais recentemente, um estudo realizado em Espanha (García-Acuña, González-Juanatey, Ezquerra, et al., 2002) junto de uma população de doentes avaliados durante um dia numa consulta de especialidade, cardiologia, ou nos cuidados de saúde primários - Estudo CARDIOTEENS - mostrou que a prevalência de FA era de 0,99% em indivíduos com menos de 50 anos, 3,90% em indivíduos com 50-59 anos, 6,32% nos 60-69 anos, 8,71% nos 70-79 anos e 11,1% em indivíduos com mais de 79 anos. Vários estudos demonstraram ainda diferenças na prevalência de FA entre homens e mulheres (Kannel & Benjamin, 1998) mais elevado junto do género masculino, mas, e como já referido, não é consensual este resultado entre os vários estudos realizados (Ferreira, 2008) tal como demonstrado no estudo de Miyasaka e colegas (2006).

Em Portugal existem poucos dados relativamente à incidência e prevalência de FA. Em 2006 foi publicado um estudo (Ascensão, 2006) realizado com recurso à Rede Sentinela (médicos de Clínica Geral e Médicos de Família que são colocados em Centros de Saúde e que de uma

forma voluntária participam semanalmente na notificação de várias situações nosológicas e estudos satélites). Tratou-se de um estudo descritivo transversal, cuja recolha de dados foi realizada entre Junho e Novembro de 2003, e no qual cada médico convidado tinha uma lista de utentes dos quais identificava os casos que tinham diagnóstico confirmado de FA. A todos os utentes identificados era aplicado um questionário desenhado para recolher informação de caracterização da patologia. Neste estudo foram incluídos 32.185 utentes, resultantes do somatório de todas as listas dos médicos participantes na rede. Foram identificados 359 casos de FA sendo estimada uma prevalência global de 0,53%. Os resultados conhecidos para homens e mulheres não apresentam variações significativas, sendo o valor de 0,53% e 0,54%, respectivamente. Por classes etárias estimou-se que a prevalência de FA era de 0,02% entre indivíduos com 35-44 anos, 0,13% dos 45-54 anos, 0,63% dos 55-64 anos, 1,83% 65-74 anos e 2,87% em indivíduos com idade igual ou superior a 75 anos.

É assumido neste estudo, que a prevalência estimada de FA para a população está submestimada nomeadamente nas classes etárias mais elevadas, justificado pelo tipo de metodologia utilizada uma vez que esta prevalência reflecte apenas um quadro de doentes seguidos pelos médicos de clínica geral e médicos de familia, não considerando outros doentes seguidos noutras especialidades. Para além disso, assume-se que a prevalência estimada encontra-se subvalorizada uma vez que não foram contabilizados os casos com doença mas não diagnosticados, isto é, não conhecidos pelo doente nem acompanhados por nenhuma especialidade médica.

1.2.1 Estudo FAMA

A falta de dados consistentes e robustos para Portugal relativamente à prevalência de FA na população Portuguesa, conduziu à realização de um estudo epidemiológico para avaliação da prevalência de Fibrilhação Auricular em Portugal sob o acrónimo de FAMA (Estudo de Avaliação da Prevalência de **F**ibrilhação **A**uricular e**m** Portug**a**l) (Bonhorst, Mendes, Adragão, et. al, 2010).

O estudo FAMA tinha como objectivo determinar a prevalência de Fibrilhação Auricular na população portuguesa com idade igual ou superior a 40 anos. A metodologia e desenho do estudo, incluindo a determinação da população do estudo e instrumentos de recolha de dados, foi desenvolvido por uma equipa de Cardiologistas especialistas na área da Arritmologia, designada por Comissão Científica do estudo.

A população deste estudo incluiu homens e mulheres residentes em Portugal (Continente e Regiões Autónomas da Madeira e dos Açores) com 40 ou mais anos de idade. Foi definida uma amostra de 10.000 indivíduos, aos quais foi aplicado um questionário de caracterização sócio-demográfica, antropométrica e clínica após obtenção do seu consentimento oral. Os dados de caracterização demográfica incluíram sexo, idade, concelho de residência, escolaridade, estado civil, situação profissional. Os dados de caracterização antropométrica consideraram o peso e altura registados com recurso a balança calibrada e através da informação constante no bilhete de identidade ou cartão do cidadão, respectivamente. Nos dados de caracterização clínica foram avaliados os conhecimentos sobre arritmias e fibrilhação auricular, factores de risco e história de doença cardiovascular no indivíduo. Foi ainda questionado se já

havia sido feito um diagnóstico médico de FA e qual o seguimento feito em casos de presença da patologia em termos de consultas e medicação tomada. Após resposta ao questionário, os indivíduos participantes neste estudo fizeram um ECG no seu domicílio, realizado por um técnico de cardiopneumologia com treino específico sobre os objectivos do estudo e procedimentos de inquirição. A realização do ECG permitiu identificar os casos de FA desconhecida e confirmar os casos identificados, através da análise dos resultados deste exame complementar de diagnóstico pela Comissão Científica do estudo.

De acordo com o protocolo deste estudo, elaborado pela Comissão Científica do estudo, para a recolha de dados foi considerada uma abordagem domiciliária, tendo sido utilizada a metodologia de *random route* para selecção dos lares. A metodologia *random route* estipulou que cada equipa de terreno tinha um ponto inicial pré-definido (por exemplo uma igreja, câmara municipal ou junta de freguesia) a partir do qual tinha um percurso aleatório parcialmente definido até à selecção do lar a abordar. No alojamento, a selecção do indivíduo também obedecia a um conjunto de critérios pré-estabelecidos por protocolo e teria que cumprir os critérios demográficos e clínicos para participação no estudo. Para além disto, era mandatório que o sujeito desse o seu consentimento oral de aceitação de participação no estudo. As avaliações dos sujeitos do estudo foram realizadas entre segunda-feira a Domingo, entre as 16h e as 21.30h, com o objectivo de captar indivíduos em diferentes situações profissionais. Os dados foram recolhidos entre Junho e Novembro de 2009. De acordo com o protocolo de estudo, a amostra foi estratificada por sexo, idade e região (NUTS II - Norte, Centro, Lisboa e Vale do Tejo - LVT, Alentejo, Algarve, Região Autónoma da Madeira - RAM, e Região Autónoma dos Açores - RAA). Esta foi definida como a primeira etapa, seguida de uma segunda etapa onde foram

seleccionados concelhos em cada uma das regiões, em número proporcional à dimensão da amostra. Os critérios de selecção dos concelhos corresponderam, por um lado, ao conhecimento empírico do território nacional e, por outro, aos valores do índice de envelhecimento da população e da densidade demográfica.

Os resultados publicados do estudo FAMA (Bonhorst, Mendes, Adragão, et. al, 2010), relativos a 10.447 indivíduos com idade igual ou superior a 40 anos, indicaram a presença de FA em 261 indivíduos considerando aqueles que já tinham diagnóstico prévio de FA e confirmado pelo ECG realizado, ou com base no exame complementar de diagnóstico realizado e analisado pela Comissão Científica do estudo. O rácio destes valores permitiu estimar um valor de prevalência de FA de 2,5% para a população portuguesa com 40 ou mais anos de idade. Foi estimado um valor total de casos de FA em Portugal, em indivíduos com idade igual ou superior a 40 anos, de aproximadamente 121.825.

A prevalência de FA pelas 7 regiões NUTS II de Portugal encontra-se descrita na tabela (1.1). Não foram reportados dados de prevalência a nível de NUTS III.

NUTS II	FIBRILHAÇÃO AURICULAR
NORTE	2,4%
CENTRO	3,5%
LVT	2,3%
ALENTEJO	2,3%
ALGARVE	1,9%
RAM	1,9%
RAA	2,0%

Tabela 1.1: Estudo FAMA: prevalência de FA por NUTS II

O estudo não identificou diferenças a nível regional e por sexo, mas em termos de idade foi consonante na evolução exponencial da FA entre classes etárias: 0,2% dos 40-49 anos, 1,0% dos 50-59 anos, 1,6% dos 60-69 anos, 6,6% dos 70-79 anos e 10,4% em indivíduos com 80 ou mais anos.

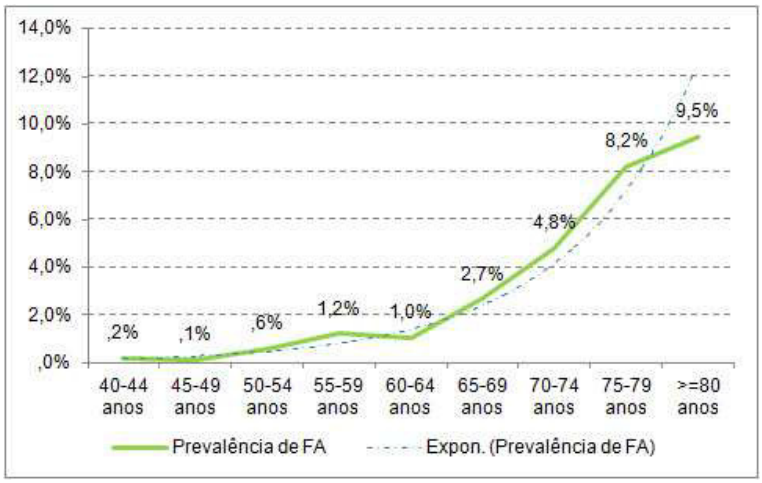


Figura 1.1: Estudo FAMA: prevalência de FA por classe etária

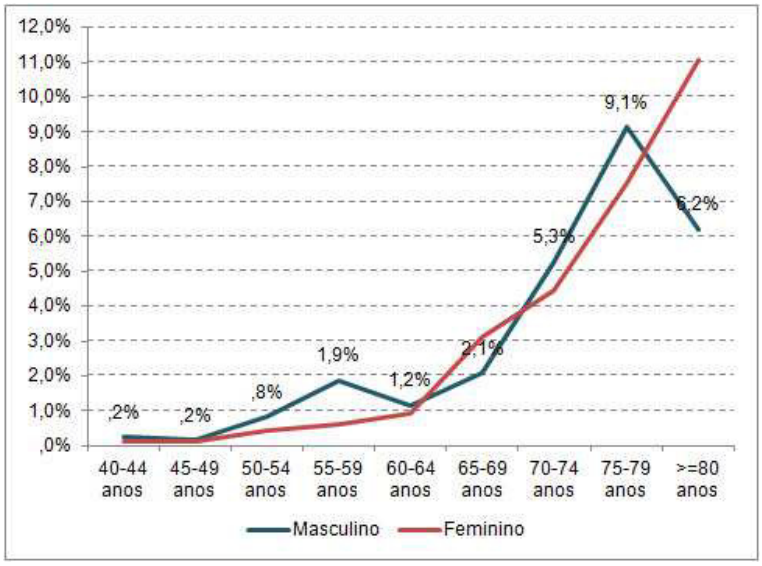


Figura 1.2: Estudo FAMA: prevalência de FA por classe etária e sexo

Outros dados de interesse publicados neste artigo incluíram uma caracterização antropomé-

trica e clínica do cluster de indivíduos com FA. De acordo com os autores (Bonhorst, Mendes, Adragão, et. al, 2010) 56% dos indivíduos com FA era do sexo feminino, a idade mediana era de 77 anos. Em termos antropométricos apurou-se que 28% destes indivíduos eram obesos e o valor médio do Índice de Massa Corporal (IMC) era de 27,7 Kg/m^2 (desvio padrão 5,3 Kg/m^2). Relativamente à prática de exercício físico, apenas 9% refere praticar de acordo com as recomendações da Organização Mundial de Saúde (OMS). Na caracterização clínica, 16% era fumador, 70% não consumia álcool, 71% sofria de hipertensão arterial, 45% tinha dislipidémia, 26% DM tipo 2, 10% hipertiroidismo e 5% doença pulmonar obstrutiva crónica (DPOC). Foi ainda referido a realização de uma análise de regressão logística para identificação de factores preditores da doença tendo sido identificados a idade, o IMC e a presença de HTA.

A discussão dos resultados destaca o facto de não terem sido encontradas diferenças na prevalência de FA entre homens e mulheres, e a estabilização dos valores de prevalência de FA no sexo masculino após os 70 anos ao invés de uma duplicação de valores. Na reflexão deste resultado foi tido em consideração informação da esperança média de vida da população portuguesa, no caso dos homens de 75 anos e 82 anos para as mulheres. Na discussão dos resultados é também reconhecido pelos autores algumas limitações associadas à metodologia do estudo e que enviesam os dados apresentados, nomeadamente o viés clínico na determinação dos casos de FA. Por viés clínico pretende-se afirmar que alguns casos de FA podem não ter sido detectados através do ECG por serem de natureza paroxística, tendo como consequência directa a subestimação dos valores de prevalência de FA. A conclusão deste estudo reforça a necessidade de dinamização de estratégias de prevenção através do controlo dos factores de risco.

1.3 WinBugs

O desenvolvimento de modelos estatísticos tem contribuído fortemente para a definição de estratégias públicas e de intervenção aos mais diversos níveis. Os avanços tecnológicos na área computacional tem permitido que os modelos estatísticos sejam cada vez mais complexos e robustos, indo de encontro ao objectivo da modelação estatística: encontrar um modelo que se ajuste aos dados, parcimonioso e interpretável. A utilização de métodos de simulação também tem contribuído em larga medida para o desenvolvimento de modelos estatísticos mais complexos, nomeadamente através da utilização do Método de Monte Carlo via Cadeias de Markov (MCMC). O MCMC constitui uma classe de algoritmos computacionais usados para simular sistemas físicos e matemáticos, e para efectuar cálculos matemáticos complexos e de difícil resolução analítica.

Em 1989, uma equipa do departamento de Bioestatística da Universidade de Oxford desenvolveu um software para análise de modelos Bayesianos com recurso a métodos de simulação de Monte Carlo via Cadeias de Markov (MCMC) e amostragem de Gibbs: *Bayesian inference Using Gibbs Sampling* (BUGS). Mais tarde, em colaboração com o *Imperial College School of Medicine* em *St. Mary's*, Londres, o *software* foi adaptado para funcionar em ambiente Windows - WinBUGS. Recentemente foi disponibilizada uma versão OpenBugs que permite correr o programa em ambiente Linux e Macintosh através do Wine (um software que permite a utilização em sistemas Unix de aplicações desenhadas para ambiente Windows) (Lawson, 2009).

O WinBugs foi desenvolvido com o intuito de resolver modelos estatísticos complexos, para os quais não era possível conceber uma solução analítica. Assumindo um modelo probabilís-

tico Bayesiano, que consiste na distribuição conjunta de toda a informação observada (dados) e não observada (parâmetros e informação omissa) para obter uma distribuição *a posteriori* para os parâmetros e informação omissa, é possível obter estatísticas empíricas e realizar inferência estatística sobre os parâmetros estimados. Neste modelo todos os parâmetros desconhecidos são tratados como variáveis aleatórias. Tal como já referido, o *software* WinBUGS utiliza MCMC para simulação e modelação dos dados. A versão clássica do programa recorria à amostragem de Gibbs para obter as distribuições *a posteriori* dos parâmetros de interesse, mas nas versões mais recentes o método de amostragem é mais geral, conferido a possibilidade de determinar o método de amostragem em função da sua distribuição (Lawson, 2009).

Capítulo 2

Objectivos de Estudo

Os resultados publicados do estudo FAMA apresentaram uma prevalência de Fibrilhação Auricular de 2,5% para a população portuguesa com 40 ou mais anos. Neste estudo não foram encontradas diferenças significativas na prevalência de FA entre homens e mulheres, ou por regiões. A prevalência foi estimada por uma metodologia clássica, ou seja, por uma abordagem frequencista onde a prevalência da doença é obtida pelo rácio entre o número de casos observados e número de casos total.

As limitações do estudo indicam que a prevalência de Fibrilhação Auricular pode estar subestimada pelo viés de memória e porque clinicamente alguns casos de FA podem não ter sido identificados no momento de realização do ECG. No entanto, outras limitações devem ser apontadas, nomeadamente sobre a metodologia de análise dos dados para estimação de valores de prevalência, uma vez que estudos realizados no âmbito do mapeamento de doenças e estimação de valores de risco referem que a utilização de uma metodologia frequencista é limitativa quanto à precisão das estimativas de prevalência quando o número de casos obser-

vados numa determinada área é relativamente baixo. Estes estudos sugerem a utilização de análise Bayesiana para a obtenção de estimativas mais precisas de valores de prevalência ou risco relativo, por região. Com o desenvolvimento da estatística computacional, a abordagem Bayesiana tem ganho relevância sobretudo pela possibilidade de considerar a variabilidade regional ou espacial ajustado a outros factores de risco, e pela utilização de métodos de simulação em modelos complexos. O WinBugs é um dos *softwares* que tem suportado e permitido o desenvolvimento de análise Bayesiana, com recurso a métodos de simulação.

Os dados publicados relativos ao estudo FAMA limitam a estimação da prevalência de FA a nível das sete regiões administrativas de Portugal (Norte, Centro, Lisboa e Vale do Tejo, Alentejo, Algarve, Região Autónoma da Madeira e Região Autónoma dos Açores - NUTS II), condicionado pelo número relativamente baixo de casos de FA observados, não identificando variações regionais na prevalência da doença. É objectivo deste trabalho fazer uso de análise Bayesiana para estimação da prevalência de FA a um segundo nível administrativo, por NUTS III, recorrendo a modelos hierárquicos Bayesianos, particularmente através do modelo desenvolvido por Besag, York e Mollié (1991).

A estrutura deste trabalho inicia com uma abordagem à doença, etiologia e factores de risco, concluída com os dados de prevalência mais relevantes a nível dos Estados Unidos da América, Europa e Portugal. No contexto dos dados de prevalência para a população portuguesa foram apresentados os dados publicados relativos ao estudo FAMA.

Após o contexto introdutório, no capítulo três, são descritos os principais conceitos associados

aos estudos epidemiológicos, tipos de estudo, efeito de confusão e processo de estandardização de dados epidemiológicos. No capítulo quatro, enquadra-se a análise Bayesiana no contexto dos estudos epidemiológicos e nos estudos de mapeamento de doença, apresentando os modelos desenvolvidos por Tsutakawa (1985), Clayton e Kaldor (1987) e Besag, York e Mollié (1991). Ainda no âmbito deste capítulo, é abordado os aspectos teóricos da convergência dos modelos na análise Bayesiana, as medidas de ajustamento, a inferência estatística e as vantagens e desvantagens da utilização destes modelos.

No capítulo cinco deste trabalho, é descrito a metodologia do estudo, que implicou uma revisão metodológica do desenho amostral do estudo FAMA. Neste capítulo é definido, no âmbito da análise Bayesiana, o modelo utilizado para estimação do valor de prevalência de FA a nível de NUTS III e respectiva parametrização. Os resultados deste trabalho são apresentados no capítulo seis, primeiro com os resultados do estudo FAMA após recálculo dos ponderadores, seguido da apresentação dos resultados da análise Bayesiana relativamente aos modelos especificados.

A discussão dos resultados, limitações e conclusões deste trabalho são apresentadas, respectivamente, nos capítulos sete, oito e nove.

Capítulo 3

Estudos Epidemiológicos

A Epidemiologia é uma ciência que tem como objecto de estudo a distribuição e identificação dos factores determinantes de uma doença numa dada população humana. Por distribuição entende-se não só a distribuição geográfica da doença mas também a distribuição da doença nas várias características da população (Gordis, 1996; Carvalho, 2007). É também do interesse da Epidemiologia a comparação de dados epidemiológicos, num dado momento ou ao longo de um determinado período de tempo, numa dada população ou mesmo entre diversas populações.

Enquanto ciência que estuda a doença, as suas características, factores condicionantes e determinantes do seu desenvolvimento, a Epidemiologia assume um papel fundamental na elaboração de políticas de saúde e implementação de medidas preventivas para a saúde pública. Independentemente do tipo de estudo epidemiológico, descritivo ou analítico, o fim último da epidemiologia é fornecer informação sobre a doença.

Os estudos epidemiológicos podem ser classificados em estudos experimentais ou de intervenção, e estudos observacionais (Gordis, 1996; Carvalho, 2007). A maioria dos estudos epidemiológicos são observacionais porque, comparativamente a um estudo experimental, são de maior facilidade de implementação, têm menores custos, maior facilidade de inclusão de sujeitos no estudo e menos considerações éticas, dado que são realizados no curso natural da doença.

Nos estudos epidemiológicos experimentais ou de intervenção, também designados por ensaios clínicos, um investigador atribui um determinado tratamento a um grupo experimental e acompanha a sua evolução ao longo do tempo de estudo relativamente ao parâmetro de avaliação. É frequente que a avaliação do grupo experimental relativamente ao parâmetro em estudo seja comparada com a de outro grupo, submetido a um outro tratamento, por vezes placebo, ou mesmo a um grupo que não foi submetido a qualquer tratamento. O número de grupos em estudo depende sempre dos objectivos de estudo.

No caso de estudos observacionais, estes podem ser classificados entre estudos descritivos e estudos analíticos. O objectivo dos estudos descritivos é determinar a distribuição da doença relativamente ao tempo, lugar e características dos indivíduos, permitindo obter respostas às questões: quando acontece, onde acontece e quem são os indivíduos que têm a doença. No caso dos estudos analíticos o principal objectivo é analisar a associação entre a doença e a exposição a esta. Existem três tipos de estudos analíticos: estudos prospectivos de coorte ou *follow-up*, estudos retrospectivos de caso-controlo, e estudos transversais ou seccionais. Os primeiros dois desenhos, ou tipos de estudo, são considerados como estudos longitudi-

nais porque são realizados ao longo de um período de tempo pré-definido, que se considera suficiente para estabelecer relações de causalidade. Por causalidade entende-se um acontecimento, condição ou característica, ou ainda uma combinação destes factores, que têm um papel fundamental no desencadear da doença. No caso do estudos de coorte ou *follow-up*, estas relações de causalidade são possíveis de serem estabelecidas existem dois grupos em estudo, um exposto e um não exposto a factores de risco identificados, que são acompanhados ao longo de um determinado período de tempo. No caso dos estudos de caso-controle, estes são longitudinais mas de natureza retrospectiva, ou seja, os dados são recolhidos relativamente a um tempo passado, para dois grupos em estudo: um com doença e outro sem doença. Deste modo, o ponto diferenciador entre um estudo de coorte e um estudo caso-controlo está nas características dos grupos de estudo, e na sua característica prospectiva ou retrospectiva. Num estudo prospectivo a característica é um antecedente (factor de risco) em que se pretende avaliar o seu efeito na população, enquanto nos estudos retrospectivos a característica é uma consequência (estádio da doença) e em que se espera compreender a causa da mesma. Existe também uma forte associação dos termos prospectivos e retrospectivos à ideia de tempo. Na realidade, a recolha dos dados nos estudos prospectivos é realizada sempre "para a frente", porque se estuda os efeitos, enquanto nos estudos retrospectivos a recolha dos dados é concretizada sempre "para trás", porque se procura a causa.

Nos estudos transversais, ou seccionais, a recolha dos dados é referente a um determinado instante no tempo, e por isso são os estudos observacionais mais simples. No entanto, ao contrário dos estudos prospectivos e retrospectivos, os estudos transversais são alvo de maior viés porque dependem directamente da memória do indivíduo e da sua capacidade de identi-

ficação à exposição de um dado factor ou à precedência de determinados sintomas.

Nos estudos epidemiológicos transversais é possível recolher dados relativos à incidência (I) de uma doença numa dada população, isto é, o número total de novos casos surgidos numa determinada população e num determinado intervalo de tempo, e dados de prevalência (P), ou seja, número total de casos existentes numa determinada população e num determinado momento temporal (Carvalho, 2007). No caso de I, a taxa de incidência é calculada pelo rácio entre o número total de novos casos observados e o número total de unidades de tempo que uma população estudada esteve em risco de contrair a doença, onde I tem uma distribuição Poisson, com parâmetro λ, em que λ representa o número de casos em risco numa determinada unidade de tempo de risco (por exemplo durante um mês, um ano ou dez anos, mas usualmente durante um ano) (Carvalho, 2007).

$$I \sim Poisson(\lambda) \tag{3.1}$$

O estimador de máxima verosimilhança para λ é dado por:

$$\hat{\lambda} = \frac{Y}{N} \tag{3.2}$$

onde $Y = y$ é o número de novos casos observados no total de pessoas-ano em risco (N). O intervalo de confiança para λ pode ser obtido através de:

$$\hat{\lambda} \pm z_{1-\frac{\alpha}{2}} \frac{\sqrt{Y}}{N} \tag{3.3}$$

onde $z_{1-\frac{\alpha}{2}}$ é quantil da distribuição normal padrão.

A taxa de prevalência é obtida através do quociente entre o número de casos identificados com a doença e o número de indivíduos pertencentes à população em estudo num determinado período de tempo. Assim:

$$A = \text{Número de casos com a doença observados} \sim Bin(n, p) \tag{3.4}$$

onde n representa o número de indivíduos na população observada e p a sua prevalência. Se o valor de n é grande tem-se que:

$$A \sim N(np, \sqrt{npq}) \tag{3.5}$$

Seja p a estimativa da taxa de prevalência de uma dada população, o intervalo de confiança para esta estimativa é dado por:

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{P(1-P)}{n}} \tag{3.6}$$

Os valores de incidência e prevalência de uma doença numa determinada população são, na maioria dos casos, diferentes embora exista uma relação de dependência dos valores de prevalência em função dos valores de incidência e da duração da doença. Uma doença crónica terá um valor de prevalência muito mais elevado do que o valor de incidência, enquanto que uma patologia com maior taxa de mortalidade após o seu diagnóstico pode apresentar um valor de incidência mais elevado comparativamente ao valor de prevalência. No caso da incidência

e a duração média da doença serem constantes ao longo do tempo, então:

$$P = I \times D \tag{3.7}$$

em que D é a duração média da doença. Em termos de aplicabilidade, a medida de prevalência é mais adequada em estudos descritivos, enquanto a medida de incidência aplica-se sobretudo em estudos etiológicos, uma vez que é uma medida menos sensível ao enviesamento causado pela sobrevivência (Carvalho, 2007).

3.1 Variáveis confundentes

Nos estudos Epidemiológicos é comum a determinação de factores de risco da doença, nomeadamente em estudos de natureza prospectiva, tal como já referido. Um dos problemas mais frequentes quando se trata de factores de risco em conjunto é a sua não independência. A possibilidade de existirem factores de risco que se encontram relacionados entre si, de tal forma que se potenciam uns aos outros, é muito elevada gerando confusão nos dados por não se compreender o seu real efeito.

Por confusão entende-se a presença de variáveis externas associadas de forma dependente e independente ao factor de exposição, ou seja, variáveis que simultaneamente são um factor de risco para a doença e estão relacionadas com o factor de exposição, alterando substancialmente a força da relação entre a doença e este (Carvalho, 2007). As variáveis de confusão são também designadas por variáveis confundentes, factor de confusão ou variável de confusão. Se num estudo epidemiológico as variáveis confundentes não forem identificadas e

controlados, as conclusões do estudo serão enviesadas e conduzirão a medidas de intervenção desapropriadas.

Existem diferentes métodos para controlar a presença de variáveis confundentes que podem ser aplicados no desenho do estudo e no decorrer da análise dos dados. Na fase de desenho é possível implementar métodos de restrição nos critérios de selecção dos participantes, ou proceder ao emparelhamento de características da população, ou ainda aleatorizar os indivíduos pelos grupos de estudo de forma a distribuir equitativamente as suas características e, consequentemente, eliminar o efeito de potenciais variáveis confundentes. Na análise dos dados é possível calcular o risco relativo bruto e comparar o mesmo com o risco relativo ajustado às potenciais variáveis confundentes. Se houver diferenças no valor de risco ajustado às variáveis confundentes, então existe evidência de confusão. Também é possível estratificar a informação e avaliar se a associação entre expostos e não expostos varia entre estratos da variável de confusão. Outra metodologia possível é através da estandardização directa ou indirecta da informação para eliminar potenciais variáveis de confusão. Ainda não fase de análise de dados é possível recorrer a análises estatísticas multivariadas, que pelas suas características permitem controlar mais do que uma variável confundente ao mesmo tempo e avaliar a magnitude do seu efeito.

3.1.1 Estandardização dos dados

A estandardização dos dados, para além de ser um método de eliminar e controlar potenciais variáveis de confusão, é uma ferramenta fundamental da epidemiologia na comparação

de valores incidência, prevalência ou taxas de mortalidade de uma doença entre diferentes populações de estudo relativamente a uma característica básica, como por exemplo o sexo, idade, nível sócio económico, que influenciam o valor de risco da doença ou morte (Natário, 2004). No processo de estandardização elimina-se a heterogeneidade das características da população e consequente variabilidade, factor que não permite que a comparação de dados epidemiológicos seja realizado de forma directa.

Nos estudos epidemiológicos a idade é uma variável elementar e a ter em consideração, uma vez que valores de incidência ou mortalidade tendem a apresentar variações significativas entre faixas etárias, nomeadamente em grupos mais velhos. A estandardização desta variável torna-se fundamental nestes estudo pelo seu potencial como factor de comparação.

O processo de estandardização pode ser feito de forma directa ou indirecta. Na estandardização directa aplica-se a taxa observada na população de interesse a uma população padrão. A escolha da população padrão depende do objectivo de comparação, isto é, se o objectivo for comparar valores com outros estudos publicados então, idealmente, deverá ser escolhida uma população externa, como por exemplo a população da Europa. No caso em que o objectivo é comparar diferentes taxas de incidência ou prevalência entre diferentes regiões de um mesmo país, deverá ser utilizada a população em estudo para estandardização dos dados, ou seja uma população interna (Natário, 2004).

No processo de estandardização directa seja (3.8) a taxa observada numa dada população de interesse i, para uma determinada classe etária j, onde Y_{ij} é o número total de casos nessa

população e N_{ij} o total da população. O número esperado de casos na população padrão é dado por:

$$\hat{\lambda}_{ij} = \frac{Y_{ij}}{N_{ij}} \tag{3.8}$$

$$E_{(i)} = \sum_j N_j^S \hat{\lambda}_{ij} \tag{3.9}$$

A taxa de estandardização directa é dada por:

$$\hat{\lambda}_i = \sum_j \frac{N_j^S}{N^S} \hat{\lambda}_{ij} \tag{3.10}$$

onde N^S representa o total na população padrão e N_j^S a dimensão de cada classe etária j na população padrão.

O erro padrão de $\hat{\lambda}_i$ é dado por:

$$SE(\hat{\lambda}_i) = \sqrt{\sum_j \left(\frac{N_j^S}{N^S}\right)^2 \frac{Y_{ij}}{N_{ij}^2}} \tag{3.11}$$

A estandardização indirecta é um dos métodos mais frequentes para estandardização de dados e relativamente fácil de implementar. Este método aplica uma taxa observada num determinado grupo etário (3.12) à dimensão da população de interesse (N_{ij}), utilizando o correspondente número observado de casos no estrato i da população de interesse. Este rácio é mais conhecido por *Standardized Mortality Ratio* (SMR) ou Taxa de Mortalidade Padronizada (3.13).

$$\lambda_j^S = \frac{Y_j^S}{N_j^S} \tag{3.12}$$

$$SMR_i = \frac{Y_i}{\sum_j N_{ij}\lambda_j^S} \tag{3.13}$$

Na SMR, $Y_i = (\sum_j Y_{ij})$ representa o total de casos na população, λ_j^S a taxa estandardizada para o grupo etário i, Y_j^S o número de casos na população de interesse para o grupo etário i e N_{ij}^S a dimensão da população de interesse.

No processo de estandardização indirecto, o número de casos esperados é dado por:

$$E_i = \sum_j N_{ij}\lambda_j^S \tag{3.14}$$

A taxa de estandardização indirecta é obtida por:

$$\hat{\lambda}_i = SMR_i \times \hat{\lambda^S} = \frac{Y_i}{E_i} \times \hat{\lambda^S} \tag{3.15}$$

3.2 Mapeamento de Doenças

Desde o estudo realizado em 1855 por John Snow sobre a epidemia da cólera (Carvalho & Natário, 2008) que a epidemiologia espacial tem ganho relevância no domínio da saúde pública, epidemiologia e estudo da doença na população humana. A epidemiologia espacial tem como objecto de estudo a descrição da doença no contexto geográfico, e as suas áreas de aplicação podem ser classificadas em análise ecológica, clusters de doenças e mapeamento de doenças. Na análise ecológica pretende-se identificar covariáveis que permitam explicar a

variação da doença no contexto geográfico, no estudo de clusters o objectivo é identificar zonas ou regiões (clusters) que confluem uma maior presença da doença em estudo, permitindo identificar a sua localização geográfica, e no mapeamento de doenças fornecer estimativas do verdadeiro risco relativo da doença numa dada área geográfica de interesse (Lawson, Browne & Rodeiro, 2003).

Os primeiros estudos desenvolvidos no âmbito do mapeamento de doenças foram realizados na área da Oncologia com um amplo trabalho desenvolvido por "The International Agency for Cancer Research"(Clayton & Kaldor, 1987) e vários trabalhos publicados por diversos autores, entre os quais destaca-se o trabalho desenvolvido por Tsutakawa (1985), Clayton e Kaldor (1987), Besag, York e Mollié (1991), Kafadar (1997), Maiti (1998), Becker (1999), Biggeri e colegas (2000), Bell e Broemeling (2000), Johnson (2004), Natário (2004), Prieto e colegas (2007) e, mais recentemente, Hegarty, Carsin e Comber (2010). A aplicação do mapeamento da doença não tem sido exclusivo na área da oncologia, como é exemplo dos estudos publicados por Marshal (1991) sobre a mortalidade infantil em Auckland, do uso de contraceptivos realizado por Borgoni e Billari (2003), a mortalidade pelo consumo de álcool de Strong e colegas (2012), ou o número de visitas de crianças e adultos devido a episódios de asma de Torabi (2013).

O mapeamento de doenças utiliza as estimativas de prevalência, incidência ou SMR de uma dada doença, obtidas a partir de contagens numa determinada área geográfica para obter as estimativas do risco relativo nessa área. Quando o número de casos observados na amostra é pequeno, a análise desta informação epidemiológica envolve um equilíbrio difícil entre

a estabilidade estatística do estimador e a precisão geográfica, dado que quando projectada esta informação num mapa a maior parte das vezes não capta a verdadeira variabilidade da população no terreno (Clayton & Kaldor, 1987), e o viés de interpretação é considerável.

Considerando uma determinada área de interesse divida em n pequenas áreas, onde $Y_i = y_i$ define o número de casos observados numa determinada área $i, i = 1...n$, e onde E_i corresponde ao número de casos esperados estimado:

$$Y_i \sim Poisson(E_i \theta i), i = 1, ..., n \tag{3.16}$$

Por máxima verosimilhança é possível estimar θ_i:

$$\hat{\theta_i} = \frac{Y_i}{E_i} \tag{3.17}$$

O erro padrão do parâmetro θ é dado por:

$$SE(\hat{\theta_i}) = \frac{1}{E_i} \tag{3.18}$$

Numa análise clássica, a projecção do risco relativo é feita através dos valores de SMR (o rácio entre o número de casos observados e número de casos esperados), e por isso é possível que grandes variações na estimativa do risco correspondam a pequenas variações no número de casos esperados, enviesando a interpretação dos resultados projectados. Para além disso, como neste caso o erro padrão é dado pelo inverso do número esperado de casos, se a estimativa para a taxa de mortalidade estandardizada for zero para áreas relativamente pequenas ou em casos de doenças raras, os valores obtidos podem também ser enviesados. Por último

a utilização das estimativas de SMR para projecção de informação epidemiológica em mapas, não permite captar a correlação espacial dos dados, existindo a uma elevada dispersão dos mesmos (*overdispersion*), o que significa que variância é mais elevada do que a esperada (Natário, 2004). Assumindo que $Y_i \sim Poisson(E_i\theta i)$, então os dados podem apresentar uma variabilidade superior àquela que o modelo de Poisson pode acomodar:

$$Var(Y_i) > E(Y_i) = E_i\theta_i \Longleftrightarrow \theta_i > 1 \tag{3.19}$$

A utilização de modelos Bayesianos no mapeamento de doenças é uma alternativa à abordagem clássica de projecção de estimativas de risco relativo, nomeadamente em situações onde amostras de pequena dimensão são fonte de imprecisão e sobredispersão.

Capítulo 4

Modelos Bayesianos

Nos estudos epidemiológicos a estimação do risco relativo e a sua inferência estatística têm sido feitas através de uma abordagem clássica, ou seja, a informação sobre um dado parâmetro θ, observado numa dada amostra y proveniente de uma população Y, é extrapolada assumindo uma dada distribuição probabilística $p(y|\theta)$, onde θ é desconhecido mas uma quantidade fixa constante (Alves, Gomes & Sousa, 2007), o que pode conduzir a interpretações enviesadas do seu intervalo de confiança (Natário, 2004).

Na abordagem Bayesiana θ é uma quantidade aleatória, com uma distribuição de probabilidade $p(\theta)$. A inferência de θ é feita com base em $p(\theta|y)$, por oposição à inferência clássica onde $p(y|\theta)$. A distribuição de probabilidade de θ é designada por distribuição *a priori*, ou seja, uma distribuição estabelecida antes da observação dos dados, actualizada posteriormente com informação obtida por uma amostra da população, ou seja, de $p(\theta)$ para $p(\theta|y)$, com base no Teorema de Bayes (4.1). A actualização da distribuição é designada por distribuição *a*

posteriori.

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta} \tag{4.1}$$

A distribuição *a posteriori* de $p(\theta|y)$ é condicional à amostra observada, e consequentemente é possível fazer inferência sobre o parâmetro θ.

A corrente clássica é crítica em relação à metodologia Bayesiana uma vez que considera que ao adoptar uma abordagem Bayesiana para inferência do parâmetro introduz-se um elemento de subjectividade e arbitrariedade na inferência estatística de θ (Greenland in Rothman, Greenland & Lash, 2008). A divergência entre estas duas correntes é forte, no entanto, nos últimos 20 anos com o desenvolvimento da estatística computacional, a abordagem Bayesiana tem sido mais utilizada, nomeadamente pela possibilidade de utilização de simulações de Monte Carlo e o Método de Monte Carlo via Cadeias de Markov (*Markov Chains Monte Carlo* - MCMC) (Natário, 2004), conceitos que abordaremos mais adiante.

Uma das vantagens reconhecidas na utilização da abordagem Bayesiana é a flexibilidade de modelização de informação e a possibilidade de incorporar nesta modelização parâmetros adicionais (Natário, 2004) como por exemplo a variabilidade espacial, elemento fundamental nos estudos de estatística espacial cujo objecto de estudo é compreender a distribuição dos dados provenientes de um determinado acontecimento, ocorrido num dado espaço.

4.1 Modelos Hierárquicos Bayesianos

Nos modelos Bayesianos todos os parâmetros são estocásticos e para os quais é definido uma distribuição *a priori*. Esta distribuição controla a forma do parâmetro e é definida pelo investigador baseado no seu conhecimento sobre o comportamento do parâmetro no contexto do fenómeno em estudo (Lawson, 2009). Nos estudos de mapeamento de doenças, é comum definir a distribuição *a priori* do parâmetro θ do modelo Poisson como uma distribuição Gama, resultando num modelo Poisson-Gama.

$$y_i|\theta \sim Poisson(e_i\theta)$$

$$\theta \sim G(\alpha, \beta)$$

Os parâmetros (α, β) podem assumir valores, ou serem consideradas variáveis aleatórias, com distribuições associadas que definem o seu comportamento. Estas distribuições são definidas como hiperprior e os parâmetros como hiperparâmetros. A noção de que os parâmetros derivam de outros, numa lógica de modelação hierárquica, é uma condição da análise Bayesiana. Os Modelos Hierárquicos Bayesianos surgem, de forma quase natural, como uma metodologia para estimação de parâmetros, sobretudo quando o número de parâmetros a estimar é elevado e estes encontram-se dependentes de parâmetros cuja estrutura não é conhecida (Natário, 2004; Lawson, 2009).

O modelo Poisson-Gama, acima definido, tem dois níveis de hierarquia: no primeiro nível θ tem uma distribuição $Gama(\alpha, \beta)$, onde α tem uma distribuição hiperprior (h_α) e β uma distribuição hiperprior (h_β) num segundo nível de hierarquia. Em termos de notação:

$$y_i|\theta \sim Poisson(e_i\theta)$$

$$\theta|\alpha,\beta \sim G(\alpha,\beta)$$

$$\alpha|\nu \sim h_\alpha$$

$$\beta|\rho \sim h_\beta$$

Nos modelos hierárquicos Bayesianos, ao contrário da abordagem clássica, não é possível apresentar estimativas pontuais do parâmetro θ uma vez que estes não são considerados como fixos mas antes derivam de um conjunto de valores de uma distribuição possível. Em função dos dados, a distribuição *a priori* do parâmetro é actualizada resultando numa distribuição *a posteriori* que caracterizará o parâmetro de interesse.

A estimação do parâmetro θ pode ser realizada considerando uma metodologia totalmente Bayesiana (*Fully Bayesian* - FB), ou através do seu estimador de máxima verosimilhança, numa abordagem Bayesiana empírica (*Empirical Bayes* - EB).

Foram vários os autores que desenvolveram modelos hierárquicos Bayesianos para estimação de parâmetros epidemiológicos, de entre os quais destaca-se os trabalhos desenvolvidos por Tsutakawa e colegas em 1985, Clayton e Kaldor em 1987, Marshall em 1990 e Besag, York e Mollié em 1991. Tsutakawa e colegas apresentaram um modelo, desenvolvido com base em dados de mortalidade provenientes de um estudo epidemiológico, que fez uso de uma aproximação integral para obter a distribuição *a posteriori*. Clayton e Kaldor sugeriram uma nova abordagem para os estimadores de SMR baseados na distribuição Poisson-Gama e log-

Normal com uma aproximação quadrática. Já na década de 90, Marshal propôs a utilização de um estimador EB sem recorrer a uma distribuição *a priori* para os valores de risco relativo, fazendo uso apenas do valor médio e variância, através do método dos momentos. Besag,York e Mollié (1991) assumiram uma posição totalmente Bayesiana (FB) para a estimação dos parâmetros recorrendo ao algoritmo de MCMC (Gomes-Rubío e López-Quílez, 2000).

Nas secções seguintes, apresentam-se os trabalhos desenvolvidos por Tsutakawa e colegas (1985), Clayton e Kaldor (1987) e Marshall (1990) no âmbito da estimação por EB, e o trabalho desenvolvido por Besag York e Mollié (1991) na perspectiva de utilização de uma abordagem FB.

4.1.1 Estimação por Empirical Bayes

Tsutakawa desenvolveu um modelo de estimação por EB motivado pelos dados de um estudo epidemiológico, desenvolvido por Marienfeld e colegas em 1980, que tinha como objectivo avaliar os efeitos do consumo de água da rede pública no Estado do Missouri nos EUA na mortalidade por cancro. Os resultados apresentados por Marienfeld e colegas demonstraram uma grande variabilidade nas taxas de mortalidade observadas por cancro do pulmão, nas diversas cidades do estado do Missouri (Tsutakawa, 1985). Breslow e Day (1975) tentaram contribuir para a explicação da variabilidade das estimativas apresentado uma nova perspectiva de análise. Estes autores consideraram que a probabilidade de morte para um dado indivíduo pode ser decomposta em dois factores, idade e região geográfica, e que estas estimativas poderiam ser calculadas pelo método de máxima verosimilhança. Contudo esta abordagem apresenta algumas limitações uma vez que a não ocorrência de morte numa dada região, com uma di-

mensão populacional pequena, conduz a uma probabilidade 0, probabilidade esta que não é considerada real (Tsutakawa, 1985).

Para lidar com a questão da variabilidade e interpretação dos dados, Tsutakawa propôs considerar o problema sob a perspectiva da modelação Bayesiana, com o desenvolvimento de um modelo de efeitos aleatórios, onde as taxas de mortalidade variam aleatoriamente de cidade para cidade de acordo com uma dada distribuição do parâmetro, e sobre a qual é possível ter informação *a priori*. Neste modelo, a frequência de ocorrência de mortes por uma determinada causa, num determinado grupo sexo-idade e num determinado período de tempo, é uma variável Poisson independente com valor médio $(\lambda_1, ..., \lambda_k)$ com $\lambda_i = n_i p_i$, onde n_i é a dimensão do grupo na região i e p_i a probabilidade de ocorrência. De acordo com o autor, p_i varia de forma aleatória e não é uma quantidade fixa, tal como identificado no trabalho de Breslow e Day (1975). Não existindo diferenças *a priori* na taxa de mortalidade entre regiões assume-se que $\theta = (\theta_1, ..., \theta_k)$, com $\theta_i = log[p_i/(1-p_i)]$, é uma variável aleatória com distribuição $\theta \sim N(\mu, \sigma)$, sendo μ desconhecido. Considera-se ainda que $\theta_i, ..., \theta_k$ são independentes. Conhecida a distribuição *a priori* é possível, pelo Teorema de Bayes, determinar a distribuição *a posteriori* do parâmetro. Dado (μ, σ), a distribuição conjunta de (θ_i, Y_i) é dada por:

$$p(\theta_i, Y_i|\mu, \sigma) = f_i(y_i|n_i, \theta_i) g(\theta_i|\mu, \sigma) \tag{4.2}$$

onde

$$f(y_i|n_i, \theta_i) = [\exp(-n_i p_i)](n_i p_i)^{yi}/y_i! \tag{4.3}$$

com

$$p_i = [1 + \exp(-\theta_i)^{-1}] \tag{4.4}$$

e

$$g(\theta_i|\mu, \sigma) = (2\pi\sigma^2)^{-\frac{1}{2}} exp[-(\theta_1 - \mu)^2/2\sigma^2] \tag{4.5}$$

A função de probabilidade marginal de Y_i é dada por:

$$p(y_i|\mu, \sigma) = \int p(y_i, \theta_i|\mu, \sigma) d\theta_i \tag{4.6}$$

A função de verosimilhança de (μ, σ), assumindo independência, é dada por:

$$\ell(\mu, \sigma) = \prod_{i=1}^{k} p(y_i|\mu, \sigma) \tag{4.7}$$

Sem perda de generalidade, para (μ, σ) conhecido, a função densidade de probabilidade (f.d.p) *a posteriori* de θ_i é dada por:

$$h(\theta_i|y_i, \mu, \sigma) = p(y_i, \theta_i|\mu, \sigma)/p(y_i|\mu, \sigma) \tag{4.8}$$

Se (μ, σ) é desconhecido, os autores propõem a sua estimação por máxima verosimilhança, através de (4.7), utilizando os valores obtido de $(\hat{\mu}, \hat{\sigma}^2)$ em (4.8). A distribuição *a posteriori* é utilizada para estimar p_i.

$$\tilde{p_i} = \int (1 + e^{-\theta_i})^{-1} h(\theta_i|y_i, \hat{\mu}, \hat{\sigma}) d\theta_i \tag{4.9}$$

Permitindo assim calcular a frequência esperada de mortes num determinado período através de:

$$\tilde{\lambda_i} = n_i \tilde{p_i}$$

Em 1987, Clayton e Kaldor desenvolveram uma nova abordagem Bayesiana para representação geográfica da incidência de cancro. De acordo com estes autores, as abordagens até então assumidas não tinham em conta a variação da população na representação gráfica dos dados, enviesando as estimativas das taxas de mortalidade estandardizada quando esta é calculada em função de poucos casos. O trabalho proposto por estes autores é crítico face a trabalhos apresentados nesta linha como o de Tsutakawa, Shoop e Marienfeld (1985).

O trabalho desenvolvido por Tsutakawa, Shoop e Marienfeld (1985) propõe que as diferenças ambientais sejam expressas para $logit(\theta_i)$ como uma combinação linear dos efeitos e interacções entre os grupos idade-sexo e as diferenças ambientais. Clayton e Kaldor apresentaram três abordagens distintas para trabalhar a questão da variabilidade espacial: uma que considera o modelo de distribuição Gama, outra que propõem a análise através da distribuição log-Normal, e a última que não recorre a distribuições paramétricas mas a modelos não paramétricos.

No primeiro modelo os autores assumem que o risco relativo (θ_i) é uma variável independente e identicamente distribuída (iid), segue uma distribuição Gama com parâmetro de escala α e forma ν, com média (ν/α) e variância (ν/α^2). O parâmetro θ_i é condicional ao número de casos observados (O_i), que segue uma distribuição Poisson, com valor esperado $\theta_i E_i$ e

variância $var(O_i) = E_i\nu/\alpha + E_i^2\nu/\alpha^2$. A distribuição *a posteriori* the θ_i, condicional a O_i, é obtida por:

$$E(\theta_i|O_i;\alpha,\nu) = \frac{O_i+\nu}{E_i+\alpha} \tag{4.10}$$

A estimativa EB para $\hat{\theta}_i$ é dada por (4.10), com as estimativas de ν e α substituídas por $\hat{\nu}$ e $\hat{\alpha}$, respectivamente. Os valores de ν e α, na distribuição binomial negativa, podem ser obtidos através do seu estimador de máxima verosimilhança, baseado na distribuição de O_i:

$$L(\alpha,\nu) = \sum_i \left[log\frac{\Gamma(O_i+\nu)}{\Gamma(\nu)} + \nu log(\alpha) - (O_i+\nu)log(E_i+\alpha) \right] \tag{4.11}$$

Definindo a primeira derivada para (4.11), relativamente a ν e α e igualando a 0, obtém-se:

$$\frac{\hat{\nu}}{\hat{\alpha}} = \frac{1}{N}\sum_i \frac{O_i+\hat{\nu}}{E_i+\hat{\alpha}} = \frac{1}{N}\sum_i \hat{\theta}_i \tag{4.12}$$

Com base em (4.12), a média da distribuição do estimador de máxima verosimilhança para θ_i é igual à média aritmética da estimativa por EB. Clayton e Kaldor sugerem que este modelo para a distribuição de θ_i pode ser estendido para a utilização de covariáveis z_i se se assumir diferentes valores para α_i, o parâmetro de escala de θ_i, e adoptando um modelo log-linear:

$$E(\theta_i) = \frac{\nu}{\alpha_i} = \exp(z_i^T\phi) \tag{4.13}$$

No modelo log-Normal, θ_i segue uma distribuição multivariada log-Normal, com β como um vector do risco relativo, com média μ e matriz de dispersão Σ. Uma vez que através desta distribuição não é possível calcular os valores de θ condicional aos valores observados O_i,

então a melhor aproximação da β é obtida por:

$$\tilde{\beta} = log\left(\frac{O_i + 1/2}{E_i}\right) \tag{4.14}$$

Os valores de β_i são iid, e a sua distribuição é definida por dois parâmetros: $\phi = \mu$ e $\Sigma = \sigma^2$. Deste modo:

$$\phi = \frac{1}{n}\sum_i b_i \tag{4.15}$$

$$b_i = \frac{\hat{\phi}' + (O_i + \frac{1}{2})(\hat{\sigma}^2)' log[(O_i + \frac{1}{2})/E_i] - (\hat{\sigma}^2)'/2}{1 + (O_i + \frac{1}{2})(\hat{\sigma}^2)'} \tag{4.16}$$

$$\hat{\alpha}^2 = \frac{1}{n}\left[(\hat{\sigma}^2)'\sum_i [1 + (\hat{\sigma}^2)'(O_i + \frac{1}{2})]^{-1} + \sum_i (b_i - \hat{\phi}')^2\right] \tag{4.17}$$

Clayton e Kaldor introduzem ainda uma variação a este modelo, assumindo que o logaritmo dos valor de risco relativo encontram-se correlacionados. Para este tipo de situação, os autores propõem a utilização de um modelo CAR (*conditional autoregression*) definido por:

$$E(\beta_i|\beta_j, j \neq i) = \mu_i + \rho\sum_j W_{ij}(\beta_j - \mu_j) \tag{4.18}$$

$$var(\beta_i|\beta_j, j \neq i) = \sigma^2 \tag{4.19}$$

Onde W é a matriz adjacente definida por:

$$W_{ij} = \begin{cases} 1, & \text{i e j são regiões adjacentes;} \\ 0, & \text{caso contrário.} \end{cases}$$

Em 1991, Marshal publicou um trabalho sobre o padrão de mortalidade na região de Auckland na Nova Zelândia. De acordo com o autor, se considerarmos uma região segmentada em N áreas indexadas a $i(i = 1, ..., N)$, com θ_i como a taxa anual de eventos na região i, e r_i o número de eventos acumulados com uma distribuição Poisson e valor médio condicional $E(r_i|\theta_i) = n_i\theta i$, onde n_i representa o número pessoas/ano. O estimador de máxima verosimilhança para o valor de θ_i é dado por $x_i = r_i/n_i$, sendo que x_i representa o valor da taxa bruta. O valor médio, condicional a θ_i, é dado por $E(x_i|\theta_i) = \theta_i$ e a sua variância obtida por $E(x_i|\theta_i) = \theta_i/n_i$.

Assumindo uma abordagem Bayesiana, e considerando que θ_i tem uma distribuição *a posteriori* com valor médio $m_i = E_\theta(\theta_i)$ e variância $A_i = var_\theta(\theta_i)$, o melhor estimador linear de Bayes, que minimiza a perda do erro quadrático total, é dado por:

$$\hat{\theta}_i = m_i + C_i(x_i - m_i) \tag{4.20}$$

onde:

$$C_i = A_i/(A_i + m_i/n_i) = var_\theta(\theta_i)/var_x(x_i) \tag{4.21}$$

De acordo com o autor, A_i e m_i pode ser estimado através do método dos momentos assumindo que, para m_i, qualquer média ponderada de x_i é um estimador não enviesado de m. A ponderação natural para esta média é dada directamente por n_i, que no seu somatório representa o número de pessoas/ano (n_i). Assim, m é estimado com base na informação amostral $\tilde{m} = \sum_i x_i n_i/n$. A_i é obtido através da variância amostral ponderada:

$$\tilde{A} = s^2 - \tilde{m}/\tilde{n} \tag{4.22}$$

Substituindo $\tilde{m}$ e $\tilde{A}$ estimados, tem-se que:

$$\tilde{\theta}_i = \tilde{m}_i + \tilde{C}_i(x_i - \tilde{m}_i) \tag{4.23}$$

onde,

$$\tilde{C}_i = \frac{s^2 - \tilde{m}/\tilde{n}}{s^2 - \tilde{m}/\tilde{n} + \tilde{m}/n_i} \tag{4.24}$$

e $\tilde{\theta}_i = \tilde{m}$ quando $s^2 < \tilde{m}/\bar{n}$.

No seu artigo, Marshall considera ainda a questão da autocorrelação espacial. De acordo com o autor, regiões vizinhas tendem a apresentar valores semelhantes na taxa de mortalidade. Assim, θ deve ser estimado, considerando os limites da vizinhança e em direcção ao seu valor médio.

4.1.2 Estimação Fully Bayes

Em 1991, Besag, York e Mollié publicaram um artigo que revê o trabalho realizado no âmbito do mapeamento de doenças e propõem uma nova metodologia de estimação do risco, numa perspectiva totalmente Bayesiana. Seja x_i o logaritmo do risco relativo de uma doença numa determinada região $i(i = 1, 2, ..., n)$, e y_i o número correspondente de casos observados de doença durante o período de estudo. Seja a patologia uma doença rara e não contagiosa, assume-se que y_i, dado x_i, é independente com distribuição Poisson e valor médio $c_i e^{xi}$, onde c_i é o número esperado de casos na zona i, e assumindo que o valor de risco é constante. Os autores adoptaram a seguinte formulação para x_i:

$$x = t + u + v \tag{4.25}$$

onde t é um termo *standard* associado a medidas de covariáveis consideradas importantes para a explicação das variações, u representa a estrutura espacial dos dados, e v representa a informação de variáveis não estruturadas. Definido o modelo e os seus parâmetros, é necessário identificar a distribuição *a posteriori* para x, condicional a u e v, assumindo que estes são independentes. Na ausência de informação, os autores assumem que v segue uma distribuição normal, enquanto u pode assumir uma distribuição da família exponencial.

$$p(u) \propto exp\left[-\sum_{j\in\partial_i} w_{ij}\phi(u_i - u_j)\right], u \in \Re^n \tag{4.26}$$

onde w_{ij} é 0, se i e j não forem regiões contíguas ou vizinhas. Assim, a densidade condicional de u é dada por:

$$p(u_i|...) \propto \exp\left[-\sum_{i<j} w_{ij}\phi(u_i - u_j)\right], u \in \Re \tag{4.27}$$

A versão mais simples de (z) proposta pelos autores é $(z) = z^2/2k$, onde k é um valor desconhecido constante positivo, e nesse caso (4.26) é actualizado para:

$$p(u|k) \propto \frac{1}{k^{n/2}} \exp\left[-\frac{1}{2k}\sum_{i\ j}(u_i - u_j)^2\right] \tag{4.28}$$

onde$i\ j$ são contíguos. A distribuição posterior de u, v, k, λ é obtida por:

$$\begin{aligned} P(u, v, k, \lambda|y) \propto & \prod_{i=1}^{n} \exp(-c_i e^{x_i})(c_i e^{x_i})^{y_i}/y_i! \\ & \times k^{-n/2} \exp -\frac{1}{2k}\sum_{i\ j}(u_i - u_j)^2 \\ & \times \lambda^{-n/2} \exp -\frac{1}{2\lambda}\sum_{i=1}^{n} v_i^2 \times prior(k, \lambda) \end{aligned} \tag{4.29}$$

onde o termo final é a densidade da prior dos dois hiperparâmetros.

As estimativas de u, v, k, λ são obtidas pela aproximação à média da distribuição posterior dada por:

$$\hat{u} = E(u|y), \hat{v} = E(v|y), \hat{k} = E(k|y), \hat{\lambda} = E(\lambda|y)$$

Estas estimativas são obtidas pelo método de amostragem de Gibbs.

Para a modelação dos dados deste modelo, os autores consideram a utilização do método de simulação de Monte Carlo via Cadeias de Markov (MCMC). Este é um método geral de simulação para obter amostras de um distribuição posterior. A simulação de amostras através do método de MCMC é um processo recursivo a partir de uma dada distribuição, razão para tal da utilização das Cadeias de Markov. Neste processo, uma sequência de variáveis aleatórias $\theta^1, \theta^2, ..., \theta^t$ estão dependentes de todas as variáveis anteriores com base na imediatamente anterior θ_{t-1}. Na sua versão mais simples:

$$\int_s g(\theta)p(\theta)d\theta \approx \frac{1}{n}\sum_{t-1}^{n} g(\theta^t) \tag{4.30}$$

onde g(.) é a função de interesse e θ^t uma amostra de $p(\theta)$.

A utilização de MCMC tem tido muito sucesso na abordagem Bayesiana uma vez que permite gerar amostras de forma arbitrária de uma densidade posterior $p(\theta|y)$ e aproximar os parâmetros de interesse. Para além disso, uma correcta implementação do algoritmo aumenta a probabilidade de convergência dos parâmetros via cadeias de Markov, aproximando-o do verdadeiro valor da distribuição em cada passo de simulação.

Em 1953, Nicholas Metropolis e colegas publicaram um artigo onde apresentaram um algoritmo de simulação que ficou conhecido como algoritmo de Metropolis Monte Carlo. Supondo que se pretende obter T amostras de uma distribuição univariada com uma função densidade de probabilidade $f(\theta|y)$. A utilização do algoritmo de Metropolis implica que exista uma valor inicial θ_0 e uma densidade proposta simétrica $q(\theta^{t+1}|\theta^t)$. No momento $(t+1)$ o algoritmo gera uma nova amostra, com base na amostra actual θ^t, sobre a qual é decidido se se aceita ou rejeita a amostra simulada. No caso de se aceitar, o algoritmo volta a gerar uma amostra repetindo todo o processo. O processo repete-se até ao ponto definido, ou seja, até ao número de amostras definido pelo investigador. Este algoritmo define uma cadeia com variações aleatórias cuja distribuição converge para a distribuição posterior esperada $f(\theta|y)$, ponto a partir do qual todas as amostras geradas são pertencentes à distribuição de interesse.

A amostragem pelo método de Gibbs é um caso particular do algoritmo de Metropolis, e à semelhança deste, foi assim designado pelo nome do seu autor Josiah W. Gibbs. Este método requer que a distribuição *a posteriori* seja decomposta numa distribuição totalmente condicional para cada parâmetro incluído no modelo e, a partir daí, obter as amostras. A amostragem pode ser eficiente se os parâmetros não apresentarem dependência entre si. Supondo $\theta = (\theta_1, ..., \theta_k)$ é um vector de parâmetros, $p(y|\theta)$ a sua verosimilhança, e $\pi(\theta)$ a distribuição *a prior*. A distribuição *a posteriori* totalmente condicional de $\pi(\theta_i|\theta_j, i \neq j, y)$ é a densidade da distribuição *a posteriori* proporcional $\pi(\theta_i|\theta_j, i \neq j, y) \propto p(y|\theta)\pi(\theta)$.

O Algoritmo de Metropolis é muito utilizado por ser pouco restritivo em relação à distribuição *a posteriori*, enquanto o algoritmo de Gibbs é mais restritivo, pois para seu uso, é necessário

conhecer as distribuições condicionais completas. Com a utilização dos métodos de simulação MCMC por amostragem de Gibbs, aos autores ultrapassam o problema computacional identificado por Marshall no seu trabalho.

Um trabalho publicado por Bernadelli e Montomoli (1992) faz uma análise critica comparativa à abordagem EB e FB considerando os modelos apresentados por Clayton e Kaldor, e o trabalho desenvolvido por Besag, York e Mollié. De acordo com os autores, a utilização de uma abordagem Bayesiana para projecção de valores de SMR é sempre mais vantajosa comparativamente à utilização do método clássico, devido à sobredispersão das estimativas. Por outro lado, considerando a modelação Bayesiana, a utilização de uma metodologia FB com recurso ao algoritmo de Gibbs deve prevalecer, uma vez que é mais potente e as estimativas obtidas são mais precisas comparativamente à abordagem EB, cujas estimativas são condicionais a um determinado ponto e, portanto, não são exactas.

4.2 Convergência dos modelos

Um dos problemas computacionais da simulação por MCMC é o momento de paragem da simulação, ou seja, o momento em que os valores dos parâmetros em estudo convergem. A análise da convergência do modelo pode ser feita de diversas formas e por diferentes métodos. Uma das abordagens mais comuns é através da análise do traçado da simulação de um conjunto de 2 a 5 cadeias paralelas, com sementes iniciais diferentes. Visualmente, uma cadeia atinge o seu ponto de estacionariedade se a distribuição dos pontos não sofrer alterações significativas com a progressão da cadeia, apresentando um valor médio e variância cons-

tante. A análise gráfica da convergência dos parâmetros apresenta alguns problemas porque é uma análise subjectiva e não existe uma garantia de que as cadeias apresentem uma dispersão estável. Como alternativa a uma medida mais subjectiva, foram desenvolvidos vários testes estatísticos que permitem analisar a convergência das cadeias de Markov, destacando-se os testes desenvolvidos por Heidelberger e Welch (1983), Geweke (1992), Gelman e Rubin (1992) e Raftery e Lewis (1992).

O teste de convergência de Heidelberger e Welch utiliza a estatística de Cramer-von-Mises para testar se um conjunto de valores amostrais é proveniente de uma distribuição estacionária. O teste é aplicado de forma sucessiva, primeiro à totalidade da cadeia, depois a um conjunto de valores que não considera 10% dos valores da cadeia e assim sucessivamente até que a hipótese nula seja aceite, ou 50% dos valores da cadeia já tenham sido descartados. No caso de se aceitar a hipótese nula, o resultado do teste indica o número de interacções necessárias para atingir a estacionariedade e a quantidade de informação não considerada. Se o resultado do teste não permitir aceitar a hipótese da estacionariedade é obtida informação para prolongar o número de interacções da cadeia via MCMC. O teste de Heidelberger e Welch apresenta ainda uma segunda componente, teste *half-width* que permite avaliar se a dimensão da amostras é adequada para estimar o valor médio com precisão. Este teste calcula um intervalo de confiança da 95% para o valor médio utilizando a parte da cadeia que passou o teste da estacionariedade. Metade do valor do intervalo de confiança estimado é comparado com o valor médio e se o rácio entre estes dois valores for inferior a um valor definido, assume-se que o valor médio estimado tem precisão, caso contrário a amostra não tem dimensão suficiente para estimar o valor com uma precisão suficiente.

O teste proposto por Geweke avalia a igualdade das médias da primeira parte da cadeia de Markov versus a parte final (por omissão os primeiros 10% e os últimos 50%). Se as amostras forem obtidas a partir de uma distribuição estacionária então as duas médias observadas são iguais. A estatística de Geweke segue uma distribuição normal estandardizada, onde a diferença entre as médias das duas amostras, assimptoticamente independentes, é dividida pela estimativa do erro padrão. Valores elevados de *z-score* indicam rejeição de estacionariedade da cadeia de Markov.

O teste desenvolvido por Raftery e Lewis permite avaliar a previsão da estimativa dos percentis estudando o número de amostras necessárias para atingir a precisão desejada da estimativa. No caso de não se atingir os valores desejados, a dimensão da cadeia de Markov deve ser superior. Este teste é apropriado para o estudo de valores de percentis.

O teste de convergência de Gelman e Rubin assenta em princípios diferentes e é um dos métodos mais populares. Os autores propõem uma abordagem mais geral à monitorização da convergência do MCMC através da análise de mais de uma cadeia de Markov. A convergência das cadeias é atingida quando os valores iniciais e finais da cadeia não se diferenciam. Para análise da convergência das cadeias realiza-se uma comparação intra e inter cadeias, muito semelhante a uma análise de variância, no qual grandes desvios nos valores das variâncias das cadeias, geralmente acima de 1, indicam uma não convergência do parâmetro e consequentemente aumento do número de simulações via MCMC. O teste de Gelman e Rubin foi desenvolvido para análise de convergência de apenas um parâmetro. Em 1997, Brooks e Gel-

man apresentaram uma versão multivariada do teste de Gelman e Rubin, permitindo a análise da convergência de múltiplos parâmetros.

4.3 Medidas de ajustamento do modelo

Conseguir a convergência de um modelo pode não ser condição suficiente para assumir que se tem um bom modelo, existindo diversos critérios que permitem avaliar a capacidade de ajustamento do modelo. Na análise Bayesiana, o BIC (*Bayesian Information Criterion*) é amplamente usado, nomeadamente em modelos hierárquicos Bayesianos. Este critério aproxima assimptoticamente um factor Bayes. Num modelo com log-verosimilhança de $\ell(\theta)$, o BIC é estimado com base nos resultados do algoritmo de MCMC através de:

$$2\hat{\ell}(\hat{\theta}) - p\ln(n) \tag{4.31}$$

onde p é o número de parâmetros lineares independentes, n o número de pontos e $\hat{\ell}(\hat{\theta})$ obtido por:

$$\hat{\ell}(\hat{\theta}) = \frac{1}{G}\sum_{i=1}^{G}\ell(\theta_i) \tag{4.32}$$

Spiegelhalter e colegas (2002) propuseram um outro critério denominado DIC (*Deviance Information Criteria*). Este critério é definido por:

$$DIC = 2E_{\theta|x}D - DE_{\theta|x}(\theta) \tag{4.33}$$

onde $D(.)$ é o desvio do modelo e x os dados observados. Este critério usa a média das amostras da distribuição posterior de θ para produzir um valor esperado de θ.

4.4 Inferência estatística

Na análise Bayesiana, dada a distribuição *a posteriori* $p(\theta|y)$, a melhor maneira de sumarizar a sua informação é através da média da distribuição *a posteriori* (4.34), do valor mediano (4.35) ou moda da distribuição *a posteriori* (4.36) (Banerjee, Carlin & Gelfand, 2004).

$$\hat{\theta} = E(\theta|y) \tag{4.34}$$

$$\hat{\theta} : \int_{-\infty}^{\hat{\theta}} p(\theta|y)d\theta = 0,5 \tag{4.35}$$

$$\hat{\theta} : p(\hat{\theta}|y) = sup_{\theta} p(\theta|y) \tag{4.36}$$

Sobre esta informação é possível realizar testes de hipóteses para inferência estatística do parâmetro. Na análise Bayesiana o teste de hipóteses tem como premissa a hipótese nula de que $H_0 : \theta \in \Theta_0$ e hipótese alternativa $H_1 : \theta \in \Theta_0^c$, onde Θ_0 é um subconjunto do espaço de parâmetros e Θ_0^c o espaço complementar. Pela distribuição *a posteriori* é possível determinar a probabilidade *a posteriori* de $P(\theta \in \Theta_0)$ e $P(\theta \in \Theta_0^c)$, ou seja, a probabilidade de H_0 e H_1 serem verdadeiras, respectivamente. Na análise Bayesiana, uma forma de aceitar a hipótese nula ou alternativa é considerando que estabelecendo um determinado limiar de probabilidade para o qual consideram que $P(\theta \in \Theta_0) > P(\theta \in \Theta_0^c)$, ou vice versa. Uma outra forma possível para aceitar a hipótese nula é considerar que a partir de um determinado limiar a hipótese é válida, ou seja $P(\theta \in \Theta_0) > 0,75$, por exemplo.

Na análise Bayesiana, o teste de hipóteses para uma estimativa pontual nula é mais difícil de

testar, sobretudo em distribuições com uma densidade contínua. O paradigma Bayesiano lida melhor com testes de hipótese para intervalos.

Relativamente ao parâmetro θ é possível determinar os seus intervalos de credibilidade. São desta forma designados porque são construídos em função da sua distribuição *a posteriori*.

Dado a distribuição *a posteriori* $p(\theta|y)$, A é o intervalo de credibilidade para θ se:

$$p(\theta \in A|y) = \int_A P(\theta|y)d\theta \tag{4.37}$$

O limite inferior e superior do intervalo de credibilidade pode ser dado pelo $100(\alpha/2)\%$ e $100(1-\alpha/2)\%$ quantil da distribuição *a posteriori*, respectivamente, se para tal considerarmos que os intervalos de credibilidade têm caudas iguais: $100(1-\alpha)\%$.

No entanto, ao contrário da noção clássica de intervalo de confiança, através dos intervalos de credibilidade é possível tecer considerações directamente sobre o parâmetro, ou seja, é possível referir que θ encontra-se, com alguma probabilidade entre os limites do intervalo de credibilidade determinado.

4.5 Vantagens e desvantagens da utilização de Modelos Bayesianos

A utilização de modelos Bayesianos traz algumas vantagens comparativamente à utilização de uma abordagem clássica ou frequencista, mas também tem algumas desvantagens. Alguns trabalhos publicados por Berger (1985), Berger e Wolpert (1988), Bernardo e Smith (1994), Carlin e Louis (2000), Robert (2001) e Wasserman (2004) descrevem as vantagens e limita-

ções da análise Bayesiana.

A utilização de informação da amostra para obter a distribuição *a posteriori* é uma das vantagens reconhecidas pelos vários autores, possibilitando a inferência estatística dos parâmetros, condicional aos dados observados e sem recorrer a uma aproximação assimptótica. Como tal, a dimensão da amostra não influencia a inferência estatística do parâmetro uma vez que esta é condicional à informação amostral. A utilização de uma distribuição *a posteriori* permite ainda que perante a disponibilização de novos dados, a distribuição *a posteriori* funcione como uma distribuição *a priori*, retirando o efeito da escolha subjectiva da distribuição *a priori*. Outra vantagem reconhecida pelos diversos autores diz respeito à interpretação do intervalo de credibilidade para a estimativa do parâmetro que, ao contrário da abordagem frequencista, é mais simples e directa.

A principal desvantagem na utilização de análise Bayesiana relaciona-se com a escolha subjectiva da distribuição *a priori*, já acima referida, e que é baseada na credibilidade do estatístico e no seu conhecimento sobre o contexto dos dados, condicionando a distribuição *a posteriori*. A abordagem Bayesiana exige também um esforço computacional adicional, sobretudo em modelos com um elevado número de parâmetros, constituindo este aspecto uma desvantagem para a utilização desta metodologia. No entanto, o desenvolvimento computacional aliado ao desenvolvimento de MCMC permitiu obviar esta questão facilitando a implementação dos modelos Bayesianos. É ainda referido como desvantagem a necessidade de utilização de sementes iniciais no processo de simulação, que na necessidade de produção de uma nova simulação, mantendo todas as restantes características, produz resultados ligeiramente

diferentes devido ao gerador de números aleatórios utilizado para a simulação. Para garantir obtenção de resultados iguais seria necessário que a semente de inicialização fosse sempre a mesma.

A utilização de modelos Bayesianos não traz vantagem acrescida à abordagem clássica ou frequencista, com a utilização de modelos paramétricos, relativamente à inferência de resultados quando a dimensão da amostra é grande. Neste caso, os resultados produzidos pelas duas abordagens são idênticos.

Capítulo 5

Metodologia

Os dados do estudo FAMA foram cedidos pelo Instituto Português do Ritmo Cardíaco (IPRC), promotor do estudo. O IPRC é uma instituição privada de solidariedade social, criada em Março de 2005 em colaboração com a Associação Portuguesa de Arritmologia, Pacing e Electrofisiologia (APAPE) e a Sociedade Portuguesa de Cardiologia (SPC), com o objectivo de promover a divulgação de aspectos clínicos, científicos e de investigação no âmbito do estudo e tratamento das arritmias cardíacas.

Tal como na maior parte dos estudos epidemiológicos mais complexos, os dados do estudo FAMA foram obtidos com recolha de informação no terreno. Foram realizadas 10.447 avaliações a indivíduos de ambos os sexos, com idade igual ou superior a 40 anos, residentes em Portugal, que contemplaram a realização de um questionário sobre dados demográficos e antropométricos, história clínica, diagnóstico prévio de FA, fontes de informação sobre a patologia, e a realização de um ECG por um técnico especializado.

Os dados disponibilizados para este projecto incluíram as principais variáveis de caracteriza-

ção demográfica e antropométrica, caracterização clínica e diagnóstico de FA.

Variáveis	**Metadados**
NUTS II	Classificação por NUTS II
NUTS III	Classificação por NUTS III
Concelho	Classificação por Concelho
Sexo	Masculino / Feminino
Idade	$\geq$40 anos
Grau de escolaridade	1 (sem escolaridade) - 13 (ensino superior)
Prática de exercício físico	Sim / Não
Hábitos tabágicos	Fumador, ex-fumador e não fumador
Consumo de álcool	Sim / Não
Hipertensão arterial	Sim / Não
Dislipidemia	Sim / Não
Diabetes	Sim / Não
Hipertiroidismo	Sim / Não
Doença Pulmonar Obstrutiva Crónica	Sim / Não
Doença Cardiovascular	Sim / Não
Diagnóstico de FA	Sim / Não

Tabela 5.1: Variáveis disponibilizadas para o projecto

Nos estudos epidemiológicos, a inferência da informação recolhida apenas junto de uma parte da população (amostra) carece do desenvolvimento de um plano amostral ou método de sondagem. Existem dois tipos de métodos de sondagem: o método probabilístico e método empírico. O método empírico, ou método de escolha judiciosa, caracteriza-se por, *a priori*, não ser possível calcular a probabilidade de um elemento vir a pertencer à amostra. Este método de sondagem é de fácil aplicação pela flexibilidade de selecção dos elementos da amostra, com custos baixos e rapidez na obtenção de informação (Costa, 2000). Os métodos de sondagens probabilísticos são mais complexos e têm como princípio base que a probabilidade de selecção de um elemento da população na amostra é conhecida e positiva (Costa, 2000). Es-

tes métodos definem um conjunto de procedimentos que permitem inferir sobre determinadas características dos parâmetros em estudo, a partir de dados provenientes de uma amostra da população. Esta inferência não se encontra livre de erro, uma vez que não se avalia todos os elementos da população, designando-se por erro de amostragem. É possível medir o grau de erro da estimativa pelo seu grau de enviesamento e precisão. O grau de enviesamento de uma estimativa é calculado pela diferença entre o valor esperado e verdadeiro valor do parâmetro (5.1):

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta \tag{5.1}$$

Para que um estimador seja considerado centrado ou não enviesado é preciso que $B(\hat{\theta}) = 0$. A determinação da precisão de um estimador θ implica a determinação da sua variância ou desvio padrão, obtidos, respectivamente, e de forma genérica pelas seguintes expressões (5.2) (5.3):

$$V(\hat{\theta}) = E[(\hat{\theta}) - E((\hat{\theta}))^2] \tag{5.2}$$

$$\sigma_{\hat{\theta}} = \sqrt{V(\hat{\theta})} \tag{5.3}$$

Assim, a precisão de um estimador é dado pelo seu Erro Quadrático Médio (EQM):

$$EQM(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = V(\hat{\theta}) + [B(\hat{\theta})]^2 \tag{5.4}$$

O plano de amostragem do estudo FAMA foi revisto e corrigido para realizar inferência para a população com base na amostra. Foi determinada a ponderação a atribuir a cada unidade amostral com base na sua probabilidade de inclusão (5.5) face ao número de Concelhos (m_h)

seleccionados em cada NUT II (M_h), e a probabilidade de selecção (5.6) de cada indivíduo na amostra considerando o número de pessoas avaliadas em cada Concelho (n_{hj}) versus a sua dimensão populacional (N_{hj}). A informação populacional foi obtida através do Instituto Nacional de Estatística (INE) relativos ao Censos de 2001 (http://www.ine.pt, acedido em 25 de Agosto de 2012).

$$\pi_{hj} = \frac{m_h}{M_h} \tag{5.5}$$

$$\pi_i h_j | h_j = \frac{n_{hj}}{N_{hj}} \tag{5.6}$$

NUTS II	**NUTS III (M_h)**	**NUTS III (m_h)**
Alentejo	46	7
Algarve	16	4
Centro	78	15
LVT	51	19
Norte	84	21
RA Açores	19	2
RA Madeira	11	2
Total	**305**	**70**

Tabela 5.2: Número de Concelhos seleccionados por NUTS II

No caso do desenho do estudo FAMA as probabilidades de inclusão são desiguais e consequentemente o peso de inclusão, ou coeficiente de extrapolação, do elemento i foi dado pelo inverso de π (Costa, 2000) (5.7):

$$w_i = \frac{1}{\pi_i} \tag{5.7}$$

Quando uma população é heterogénea deve considerar-se a utilização de um método de sondagem probabilístico estratificado (Costa, 2000). Este método utiliza informação auxiliar para

identificar e dividir a população em estratos (H), ou sub-grupos homogéneos, mutuamente exclusivos, com dimensão $N_1...,N_h...,N_H$. Em cada estrato é seleccionado de forma independente uma dimensão amostral $n_1...,n_h...,n_H$, tal que $\sum n_h = n$ resulta na dimensão da amostra fixa *a priori*.

Uma amostra aleatória estratificada apresenta algumas vantagens comparativamente ao uso de uma amostra aleatória simples, porque quando realizada correctamente, apresenta estimativas mais precisas quer para a população total, quer para os subgrupos identificados relativamente ao parâmetro de interesse. O número de estratos identificados depende da informação auxiliar disponível.

Quando a estratificação da amostra não é obtida através do desenho de estudo é possível proceder-se a uma processo de pós-estratificação. Isto significa que a estratificação é feita depois da amostra ter sido recolhida, assumindo, tal como na sondagem aleatória estratifica, a existência de sub-grupos homogéneos na população em estudo (Costa, 2000). A metodologia de pós-estratificação também necessita de informação auxiliar para definição dos estratos, que devem ser o mais homogéneos possível. Na pós-estratificação assume-se que as dimensões dos pós-estratos são conhecidas e, como tal, os pesos iniciais definidos são ajustados de modo a que a distribuição da amostra seja novamente ponderada pelos valores da população, e esteja mais de acordo com a distribuição da população pelos vários estratos definidos.

A pós estratificação apresenta algumas vantagens uma vez que permite capturar uma grande parte da variância do estimador introduzida pelo desenho do estudo, se os pós-estratos forem

homogéneos. A possibilidade de adaptação deste método ao tratamento de valores extremos ou *outliers*, desde que seja possível ponderar os estratos que contêm estes indivíduos, é outra das vantagens referidas para utilização da pós-estratificação (Costa, 2000).

Na pós-estratificação o estimador do total da população é dado, genericamente, por:

$$\hat{\tau}_{PS} = \sum_{i=1}^{L} N_i \frac{\hat{\tau}_i}{\hat{N}_i} \tag{5.8}$$

onde $\hat{\tau}_i$ e $\hat{N}_i$ são os estimadores centrados do total e da dimensão do *i-ésimo* pós estrato, respectivamente, obtidos através de:

$$\hat{\tau}_i = \hat{\tau}_{i\pi} = \sum_{k \in S_i} \frac{Y_k}{\pi_k} \tag{5.9}$$

O estimador do valor médio é dado por:

$$\hat{\mu}_{PS} = \frac{1}{N} \hat{\tau}_{PS} \tag{5.10}$$

Os dados do estudo FAMA foram pós-estratificados para correcção do desenho do estudo, assumindo a presença de estratos ao nível de NUTS II, sexo e classes etárias. A informação auxiliar utilizada no processo de pós-estratificação foi obtida através do Instituto Nacional de Estatística, no Censos de 2001, por ser a informação oficial da população portuguesa à data da recolha dos dados do estudo FAMA (http://www.ine.pt, acedido em 25 de Agosto de 2012).

A pós-estratificação dos dados do estudo FAMA foi realizada em R, versão R 2.12.1 (2010-12-16) com recurso ao *package survey*.

5.1 Análise Bayesiana

Com os dados pós-estratificados, foi determinado o número observado de casos de FA por região (O_i) e, por estandardização indirecta, o número esperado de casos (E_i), que permitiu calcular por máxima verosimilhança o valor de *Standardized Disease Ratio* (SDR), por região (NUTS III), equivalente à estimação de SMR:

$$SDR = \frac{O_i}{E_i}$$

A estes dados foi aplicado o modelo proposto por Besag, York e Mollié (1991) para estimação do risco relativo de FA por NUTS III. Este é um modelo Poisson, onde o número observado de casos para cada estrato ou região é uma variável dependente do número esperado de casos, do valor de risco inicial e de uma componente relativa à estrutura espacial dos dados. O modelo é definido da seguinte forma:

$$O_i \sim Poisson(\mu_i)$$

$$log(\mu) = log(E_i) + \alpha_0 + b_i$$

onde α_0 representa o valor inicial de risco relativo entre as regiões em estudo, e b_i um efeito aleatório que capta os valores residuais ou latentes do risco relativo da doença na região i e que, neste caso, define também uma dependência espacial dos dados. O parâmetro α_0 é modelado a partir de uma distribuição Uniforme:

$$\alpha_0 \sim Uniforme()$$

A componente da estrutura espacial dos dados é modelado por um modelo condicional auto-regressivo (CAR), com a definição de uma matriz referente às regiões adjacentes ou contíguas (W), e onde para regiões adjacentes ou contíguas $W_{ij} = 1$, caso contrário $W_{ij} = 0$. O modelo CAR segue uma distribuição Normal com média condicional ao número médio de regiões vizinhas, e variância condicional inversamente proporcional ao número de regiões vizinhas. O modelo CAR é definido considerando um vector com o número adjacente de regiões em cada área [adj], o vector W, um vector com a identificação das regiões adjacentes em cada região [num], e um argumento escalar relativo ao inverso da variância e que também representa um argumento de precisão (τ).

$$S[1:N] \sim CAR.Normal(adj[], W[], num[], \tau)$$

O argumento τ é modelado a partir de uma distribuição Gama:

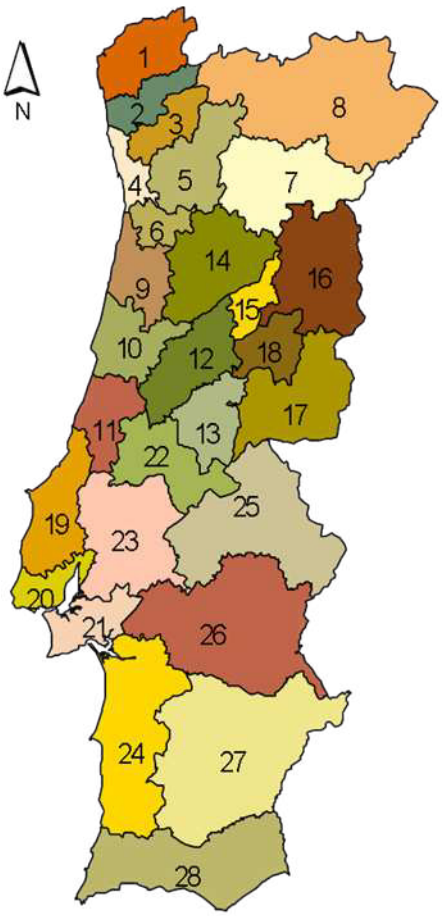
$$\tau \sim \ Gamma(\alpha, \beta)$$

A matriz de regiões adjacentes a utilizar na modelação dos dados do estudo FAMA foi estruturada em função de NUTS III e encontra-se descrita na figura (5.1).

O modelo de BYM foi adaptado para a introdução de covariáveis, definido genericamente da seguinte forma:

$$log(\mu) = logE_i + \alpha_0 + \alpha_{1,\ldots,n} X_{1,\ldots,n} + b_i$$

onde $X_{1,\ldots,n}$ representa as covariáveis e $\alpha_{1,\ldots,n}$ os respectivos coeficientes de ponderação. Os coeficientes de ponderação das covariáveis são modelados a partir de uma distribuição



(1) Minho Lima: 2,8,3,4
(2) Cávado: 1,3,8,4
(3) Ave: 2,4,5,8,1
(4) Grande Porto: 2,3,5,6,9,1
(5) Tâmega: 3,4,6,7,8,13
(6) Entre Douro e Vouga: 4,5,7,13,9
(7) Douro: 6,8,5,13,17
(8) Alto Trás os Montes: 1,2,3,5,7
(9) Baixo Vouga: 4,6,13,10,12,
(10) Baixo Mondego: 9,13,12,11
(11) Pinhal Litoral: 10,12,22,23,19
(12) Pinhal Interior Norte: 10,9,13,15,16,14,22,11,18
(13) Pinhal Interior Sul: 10,9,6,5,7,17,15,12
(14) Dão Lafões: 22,12,16,18,25
(15) Serra da Estrela: 13,17,16,12
(16) Beira Interior Norte: 18,14,12,15,17
(17) Beira Interior Sul: 7,13,15,16,18
(18) Cova da Beira: 25,14,17,16,12
(19) Oeste: 11,23,20
(20) Grande Lisboa: 19,23,21
(21) Península de Setúbal: 20,23,26,24
(22) Médio Tejo: 25,23,11,12,14
(23) Leziria do Tejo: 22,11,19,20,21,26,25
(24) Alentejo Litoral: 21,26,27,28
(25) Alto Alentejo: 18,14,22,23,26
(26) Alentejo Central: 25,23,21,24,27
(27) Baixo Alentejo: 26,24,28
(28) Algarve: 24,27

Figura 5.1: Matriz de regiões adjacentes por NUTS II

Normal:

$$\alpha_{1,\ldots,n} \sim Normal(\mu, \sigma)$$

As covariáveis identificadas para inclusão no modelo de BYM derivaram de uma análise prévia de identificação dos factores de risco. A idade e o valor do Índice de Massa Corporal foram identificadas como potenciais covariáveis a ajustar no modelo. A variável idade foi reclassificada sob a percentagem de indivíduos com idade acima dos 65 anos, por ser uma idade relevante na variação da prevalência de FA na população Portuguesa. Ambas variáveis foram estandardizadas para inclusão no modelo. Não foi possível incluir no modelo informação rela-

tiva aos factores de risco clínico por não ter sido observada convergência dos parâmetros no modelo proposto por BYM aquando do seu estudo.

Com recurso ao WinBugs, o método de MCMC foi utilizado para obter uma amostra da distribuição *a posteriori* dos parâmetros do modelo, gerando automaticamente amostras da distribuição marginal *posteriori* e dos hiperparâmetros. A amostragem de Gibbs foi realizada para 100.000 interacções, descartando as 10.000 primeiras interacções como recomendado (Dellaportas, Forster & Ntzoufras, 2002). Na simulação de amostras foram definidas duas cadeias. A convergência da distribuição *a posteriori* foi avaliada pelos seus parâmetros através de diagramas de dispersão, correlogramas e da estatística de Gelman-Rubin implementada no *software* utilizado. O ajustamento dos modelos estimados foi avaliado pelo *Deviance Information Criteria* (DIC) e o número de parâmetros (pD). Foi realizada uma análise de sensibilidade para os modelos definidos, alterando o valor das priors iniciais.

O código computacional para ambos modelos, dados e valores iniciais, encontra-se no anexo 2 deste trabalho.

Capítulo 6

Resultados

6.1 Resultados do Estudo FAMA após recálculo dos ponderadores

6.1.1 Total de casos de FA

O número total de casos estimados de FA na população portuguesa com 40 anos ou mais foi de 138.427 (erro padrão 9.181 casos). As estimativas para o total de casos de FA por NUTS II encontram-se descritos na tabela (6.1).

NUTS II	Total de casos de FA	Erro Padrão
NORTE	42.353	5.085
CENTRO	34.589	6.033
LVT	45.262	4.162
ALENTEJO	7.452	1.699
ALGARVE	2.886	813
RAM	3.792	934
RAA	2.090	534

Tabela 6.1: Estimativa do total de casos de FA por NUTS II

A estimativa do total de casos com FA por sexo revela que o número de casos no sexo feminino é 1,4 vezes superior ao estimado para o sexo masculino. O número de casos estimados para o sexo feminino é de 81.437 casos (erro padrão 5.245 casos) e no sexo masculino de 56.989 casos (erro padrão 7.147 casos).

SEXO	Total de casos de FA	Erro Padrão
Feminino	81.437	5.245
Masculino	56.989	7.147

Tabela 6.2: Estimativa do total de casos de FA por sexo

Relativamente à idade, a estimativa total de casos de FA por classe etária encontra-se descrita na tabela (6.3). Observa-se que o número de casos aumenta com a progressão da idade, nomeadamente a partir dos 65 anos de idade.

Classes Etárias	Total	Erro Padrão
40-44 anos	1.412	783
45-49 anos	1.170	827
50-55 anos	4.430	1.500
55-59 anos	8.351	2.331
60-64 anos	6.403	1.576
65-69 anos	13.795	2.822
70-74 anos	22.974	3.290
75-79 anos	34.687	4.488
80-84 anos	24.546	3.988
85+ anos	20.655	3.091

Tabela 6.3: Estimativa do total de casos de FA por classe etária

O número total de casos de FA foi também estimado para o valor de IMC (6.4) e hábitos tabágicos (6.5). O número de casos de FA é superior em indivíduos com excesso de peso ou obesidade (IMC com valor superior a 25 kg/m2), e em indivíduos fumadores ou ex-fumadores, mas particularmente nos fumadores.

IMC	Total	Erro Padrão
<18,5 kg/m2	2.651	1.389
18,5-25,0 kg/m2	43.029	5.448
25,1-30,0 kg/m2	50.693	5.953
>30,0 kg/m2	39.833	3.845

Tabela 6.4: Estimativa do total de casos de FA por classe de IMC

IMC	Total	Erro Padrão
Não fumador	4.179	1.210
Ex-fumador	26.214	4.093
Fumador	107.339	8.349

Tabela 6.5: Estimativa do total de casos de FA por hábitos tabágicos

6.1.2 Prevalência de FA

A prevalência de FA, obtida através do valor médio do parâmetro, foi estimado para o total, por região (NUTS II), sexo, classes etárias, hábitos tabágicos, classes de IMC e presença de factores de risco. Globalmente, a prevalência estimada de FA foi 2,52% (erro padrão 0,17%), com um Intervalo de Confiança a 95% de [2,19%; 2,85%]. A prevalência de FA por NUTS II evidenciou valores superiores ao valor global na região Centro, no Alentejo e na Região Autónoma da Madeira (RAM). A prevalência de FA é inferior ao valor médio nacional na região do Algarve e na Região Autónoma dos Açores. As diferenças encontradas são estatisticamente significativas (p-value=0,0199 para um erro α=0,005).

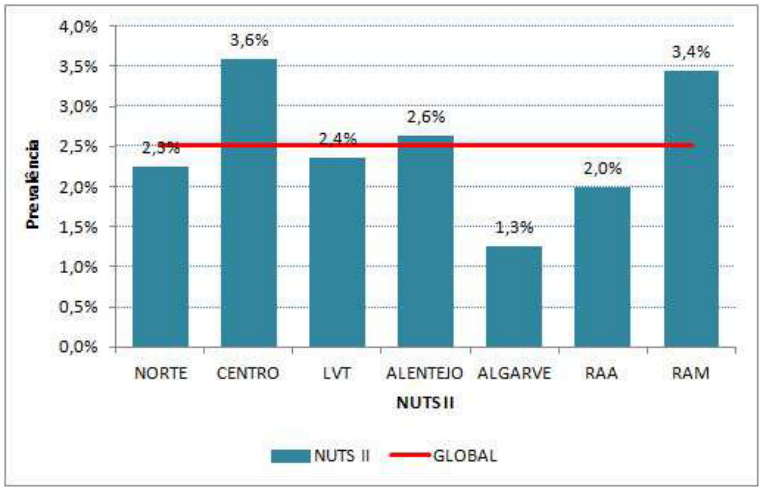


Figura 6.1: Prevalência de FA por NUTS II

A prevalência de FA no sexo feminino, na população portuguesa com 40 ou mais anos de idade, foi de 2,75% e no sexo masculino de 2,25%. As diferenças observadas na prevalência de FA entre estes dois grupos não releva ser estatisticamente diferente (p-value=0,1037 para para um erro α=0,005).

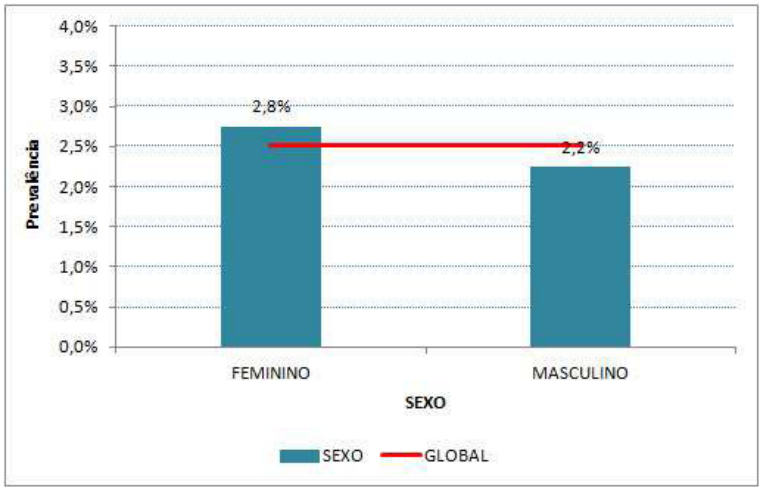


Figura 6.2: Prevalência de FA por sexo

A prevalência de FA por classes etárias aumenta com a idade, como já referido anteriormente. A partir dos 65 anos de idade observa-se um aumento exponencial no valor médio de casos de FA, duplicando o valor por cada cinco anos até à idade de 80 anos. A partir dessa idade o número de casos mantém um ritmo de crescimento, contudo mais lento. A prevalência de FA por classes etárias apresenta diferenças estatisticamente significativas (p-value<0,0001 para um erro α=0,005).

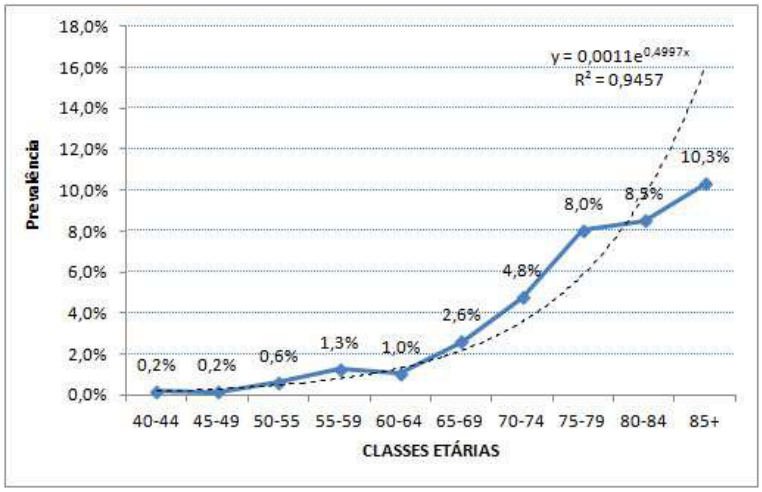


Figura 6.3: Prevalência de FA por classes etárias

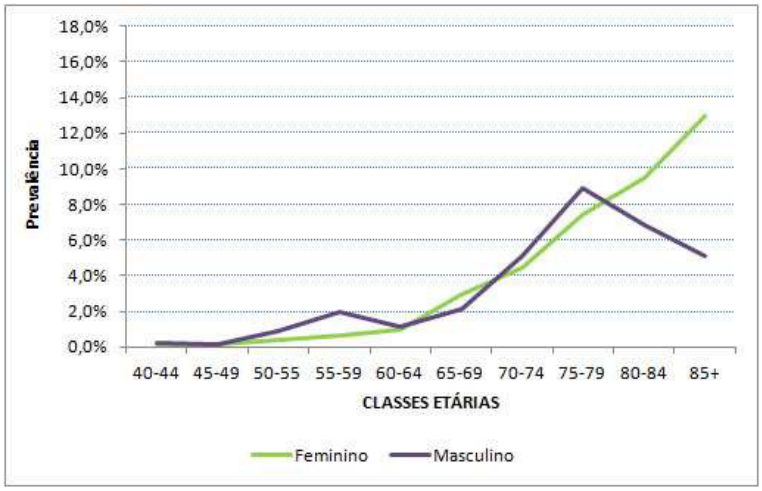


Figura 6.4: Prevalência de FA por classes etárias e sexo

Na análise de FA por classes de IMC, observa-se que nas classes mais extremas a prevalência de FA é superior à média global. As diferenças observadas são estatisticamente significativas (p-value=0,0022 para para um erro α=0,005).

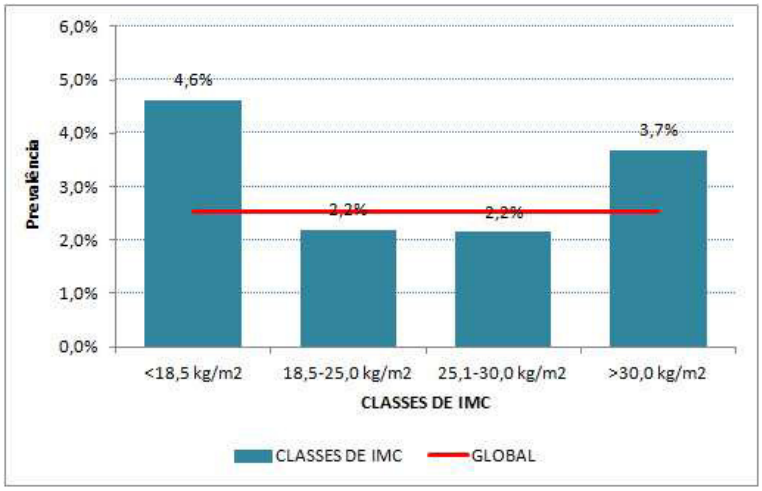


Figura 6.5: Prevalência de FA por classes de IMC

Relativamente aos hábitos tabágicos a prevalência de FA é superior em indivíduos fumadores (2,97%) e ex-fumadores (2,74%) comparativamente a indivíduos que não têm hábitos tabágicos (0,48%). As diferenças observadas são estatisticamente significativas (p-value<0,0001

para para um erro α=0,005).

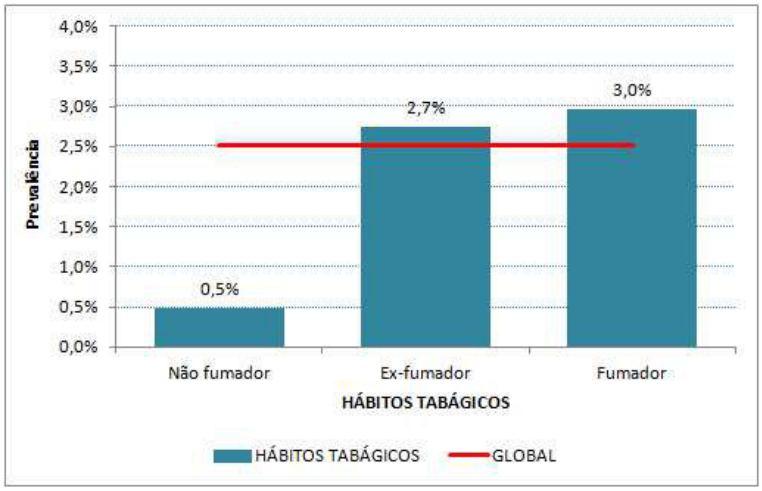


Figura 6.6: Prevalência de FA por hábitos tabágicos

Ao nível dos factores de risco, HTA, Dislipidemia, Diabetes, DPOC, Hipertiroidismo, AVC e História familiar de DCV, a sua presença é uma factor para que a prevalência de FA seja mais elevada comparativamente ao valor global.

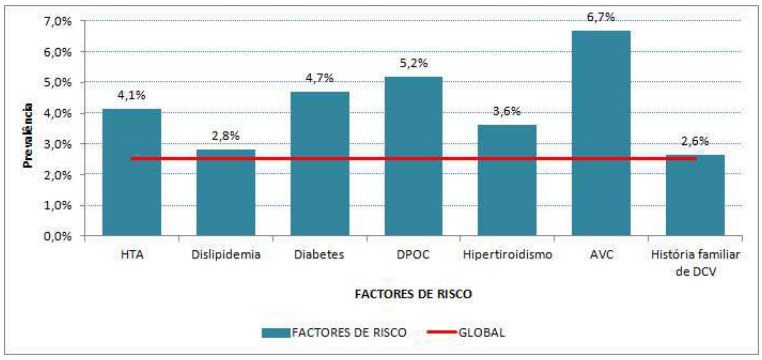


Figura 6.7: Prevalência de FA por factores de riscos

6.1.3 Factores de Risco para a FA

Foi realizado um modelo linear generalizado para identificar os factores de risco para a FA. As variáveis consideradas foram a idade, classificada em indivíduos com idade igual ou superior a 65 anos e idade inferior a 65 anos, hábitos tabágicos (ser fumador ou não), valor de índice de

massa corporal (IMC), e todos os factores de risco para a doença (HTA, Dislipidemia, Diabetes, DPOC, Hipertiroidismo, AVC e História familiar de DCV).

Os resultados indicam que a idade, os hábitos tabágicos, o valor de IMC, a presença de HTA, e a presença de DPOC são os factores que tem impacto significativo na variável dependente, ou seja, a prevalência de FA. As variáveis com impacto significativo para a FA foram consideradas num modelo de análise Bayesiana.

Variável	Coeficientes	Erro padrão	t value	p-value
(Intercept)	-7,36325	0,50531	-14,572	<0,001
Idade	2,12605	0,22798	9,326	<0,001
Fumador	0,95370	0,32262	2,956	0,00467
IMC	0,04070	0,01572	2,590	0,01243
FR: HTA	0,44367	0,15657	2,834	0,00653
FR: História familiar de DCV	0,24447	0,14094	1,735	0,08875
FR: Dislipidémia	-0,34391	0,13955	-2,464	0,01706
FR: Diabetes	0,03105	0,17943	0,173	0,86329
FR: Hipertiroidismo	0,14420	0,29334	0,492	0,62508
FR: DPOC	0,66049	0,28545	2,314	0,02466
FR: AVC	0,37599	0,20689	1,817	0,07493
FR: Enfarte	0,29882	0,25656	1,165	0,24945

Tabela 6.6: Resultados do Modelo Linear Generalizado

6.2 Resultados Análise Bayesiana

A estimação dos valores de SDR encontra-se descrita na tabela (6.7). Na mesma tabela encontram-se os valores do número de casos observados de FA NUTS III (O_i) e o valor esperado (E_i). As estimativas de SDR por NUTS III variam entre 0,26, observado na região do Baixo Alentejo, e um valor de 6,95 estimado para a região da Serra da Estrela. A figura (6.8)

é a representação num mapa de coropletos das estimativas de SDR, por NUTS III. A falta de observações em algumas regiões não permitiu estimar o valor de SDR para essas mesmas regiões, designadamente a região do Cávado, Ave, Entre Douro e Vouga, e Alto Trás-os-Montes, Beira Interior Norte, Beira Interior Sul e Cova da Beira na região Centro, e a Península de Setúbal na região de Lisboa. Estas regiões encontram-se identificadas a preto no mapa (6.8).

	NUTS III	O_i	E_i	SDR_i
RR[1]	Minho-Lima	1826	4280	0,427
RR[2]	Cávado	-	4901	-
RR[3]	Ave	-	6372	-
RR[4]	Grande Porto	35233	17838	1,975
RR[5]	Tâmega	1842	6410	0,287
RR[6]	Entre Douro e Vouga	-	3855	-
RR[7]	Douro	3452	3438	1,004
RR[8]	Alto Trás-os-Montes	-	4095	-
RR[9]	Baixo Vouga	9643	5980	1,613
RR[10]	Baixo Mondego	1951	5619	0,347
RR[11]	Pinhal Litoral	1328	4145	0,320
RR[12]	Pinhal Interior Norte	2290	2576	0,889
RR[13]	Dão-Lafões	11906	4900	2,430
RR[14]	Pinhal Interior Sul	1344	910	1,477
RR[15]	Serra da Estrela	6128	882	6,955
RR[16]	Beira Interior Norte	-	2120	-
RR[17]	Beira Interior Sul	-	1531	-
RR[18]	Cova da Beira	-	1647	-
RR[19]	Oeste	4221	5769	0,732
RR[20]	Médio Tejo	1950	4038	0,483
RR[21]	Grande Lisboa	35207	30615	1,150
RR[22]	Península de Setúbal	-	11552	-
RR[23]	Alentejo Litoral	2302	1802	1,277
RR[24]	Alto Alentejo	720	2320	0,310
RR[25]	Alentejo Central	3831	3182	1,204
RR[26]	Baixo Alentejo	600	2302	0,261
RR[27]	Lezíria do Tejo	3885	4280	0,908
RR[28]	Algarve	2886	6840	0,422
RR[29]	Região Autónoma dos Açores	2091	2697	0,788
RR[30]	Região Autónoma da Madeira	3792	2867	1,322

Tabela 6.7: SMR por NUTS III

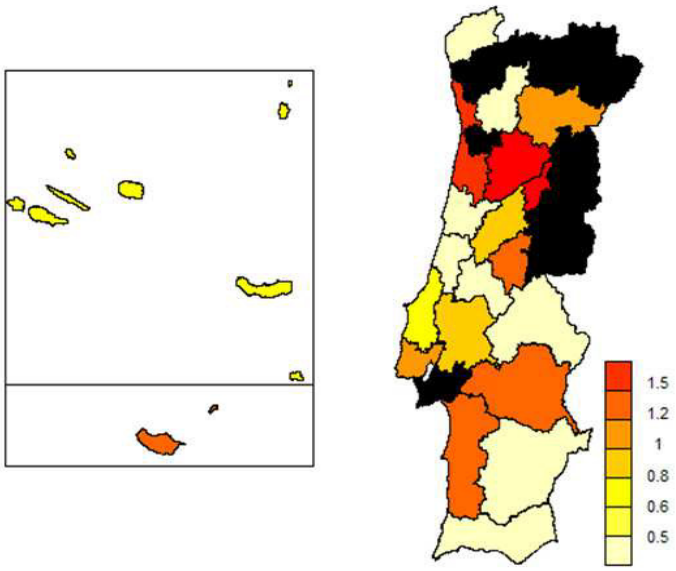


Figura 6.8: SMR por NUTS III

Os resultados do modelo de BYM sem a introdução de covariáveis (Modelo I) e com a modelização incluindo a variável idade (Modelo II) e IMC (Modelo III) encontram-se descritos na tabela (6.8). Esta tabela sumariza as medidas da distribuição marginal posterior dos parâmetros de interesse, α_0 (ponto de intersecção) e τ para o modelo I, α_0, α_1 (coeficiente associado à covariável) e τ para o modelo II e II respectivamente, obtidos por MCMC, com os correspondentes intervalos de credibilidade. Observa-se uma convergência dos parâmetros de interesse, com o valor da estatística de Gelman-Rubin próximo de 1. No modelo I, sem a introdução de covariáveis, o parâmetro α_0 não apresenta um valor significativo, mas o τ é significativo indicando a existência de uma componente associada à estrutura espacial dos dados. No modelo II, o parâmetro da idade embora significativo, tem um coeficiente de ponderação relativamente baixo e negativo. Neste modelo, o coeficiente associado à estrutura espacial dos dados continua a significativo. Para o modelo III a variável IMC também apresenta um coeficiente de ponderação significativo, positivo, mas relativamente baixo.

a priori			MCMC *a posteriori*					
Modelo	Parâmetro	Distribuição	Média (DP)	Média (DP)	2,5%	Mediana	97,5%	Gelman-Rubin
I	α_0	U()	-	-0,11 (0,10)	-0,31	-0,11	0,08	0,999
	τ	G(0,5; 0,05)	10,0 (14,1)	0,34 (0,11)	0,16	0,33	0,58	0,999
II	α_0	U()	-	-0,11 (0,10)	-0,30	-0,10	0,10	0,999
	α_1	N(0; 0,0005)	0 (0,02)	-0,10 (0,02)	-0,14	-0,10	-0,06	0,999
	τ	G(0,5; 0,05)	10,0 (14,1)	0,33 (0,10)	0,16	0,33	0,56	0,999
III	α_0	U()	-	-0,10 (0,10)	-0,30	-0,10	0,10	1,00
	α_1	N(0; 0,0005)	0 (0,02)	0,12 (0,03)	0,06	0,12	0,17	1,00
	τ	G(0,5; 0,05)	10,0 (14,1)	0,32 (0,10)	0,15	0,31	0,54	1,00

Tabela 6.8: Resultados da amostragem de Gibbs

A análise dos valores de *deviance* dos modelos estudados (6.9) mostra que a introdução das variáveis idade (relativa à percentagem de indivíduos com 65 ou mais anos) e o valor médio de IMC melhora a performance do modelo, com um menor valor comparativamente ao modelo sem covariáveis. Contudo, a modelação destas duas covariáveis em conjunto não é uma mais valia considerando o valor de *deviance* obtido. Entre o modelo II e III, o modelo que inclui o valor médio de IMC aponta ser uma melhor opção, com um valor de *deviance* relativamente mais baixo e com menor uso de parâmetros.

	DIC	pD	*Deviance*
Sem covariáveis	4.922,38	2.069,60	2.852,78
Covariáveis (Idade)	4.919,77	2.151,48	2.768,29
Covariáveis (IMC)	4.831,23	2.120,20	2.711,03
Covariáveis (Idade e IMC)	16.861,10	8.319,36	8.541,74

Tabela 6.9: *Deviance information criterion* para os modelos estimados

A análise de sensibilidade realizada, na qual foram alterados os valores iniciais das priors e aumentando o valor da variância, mostrou uma consistência nos valores da distribuição *a*

posteriori dos parâmetros estimados (6.10)

		a priori		MCMC *a posteriori*				
Modelo	Parâmetro	Distribuição	Média (DP)	Média (DP)	2,5%	Mediana	97,5%	Gelman-Rubin
I	α_0	U()	-	-0,11 (0,10)	-0,31	-0,11	0,09	0,999
	τ	G(0,05; 0,005)	10,0 (44,7)	0,33 (0,11)	0,15	0,31	0,56	0,999
II	α_0	U()	-	-0,11 (0,09)	-0,31	-0,11	0,08	0,999
	α_1	N(0; 0,005)	0 (0,07)	-0,09 (0,03)	-0,14	-0,09	-0,03	0,999
	τ	G(0,05; 0,005)	10,0 (44,7)	0,35 (0,11)	0,17	0,33	0,59	0,999
III	α_0	U()	-	-0,10 (0,10)	-0,31	-0,10	0,11	1,00
	α_1	N(0; 0,005)	0 (0,07)	0,12 (0,03)	0,06	0,12	0,17	1,00
	τ	G(0,05; 0,005)	10,0 (44,7)	0,30 (0,10)	0,14	0,29	0,53	1,00

Tabela 6.10: Resultados da amostragem de Gibbs - análise de sensibilidade

Os valores de Risco Relativo (RR) estimados através do modelo de BYM encontram-se descritos na tabela (6.11). As regiões do Baixo Alentejo e Serra da Estrela são as regiões onde o valor de risco relativo é menor (0,27) e maior (7,05), respectivamente. Devido à componente espacial do modelo de BYM e da utilização da matriz de regiões adjacentes, foi possível obter estimativas de risco relativo para regiões onde não foi recolhida informação, nomeadamente na região do Cávado, Ave, Entre Douro e Vouga, e Alto Trás-os-Montes na região Norte, Beira Interior Norte, Beira Interior Sul e Cova da Beira na região Centro, e a Península de Setúbal na região de Lisboa. No entanto, para as estimativas apresentadas observa-se que o desvio padrão da distribuição *a posteriori* é mais elevado nestas regiões, para as quais não existia informação disponível, comparativamente às restantes regiões.

				Modelo sem Covariáveis		Modelo com Covariáveis: Idade		Modelo com Covariáveis: IMC		Regiões Adjacentes
NUTS III	O[i]	E[i]	SMR[i]	Média (DP)	IC 95%	Média (DP)	IC 95%	Média (DP)	IC 95%	
RR[1] Minho-Lima	1826	4280	0,43	0,43 (0,01)	[0,41 ; 0,45]	0,43 (0,01)	[0,41; 0,45]	0,43 (0,01)	[0,41; 0,45]	2,8,3,4
RR[2] Cávado	NA	4901	-	1,20 (1,70)	[0,12 ; 4,91]	1,33 (2,01)	[0,12; 5,53]	1,29 (3,21)	[0,11; 5,43]	1,3,8,4
RR[3] Ave	NA	6372	-	0,93 (1,09)	[0,12 ; 3,40]	1,02 (1,19)	[0,13; 3,79]	1,03 (1,30)	[0,13; 3,86]	2,4,5,8,1
RR[4] Grande Porto	35233	17838	1,98	2,00 (0,25)	[1,56 ; 2,54]	1,99 (0,25)	[1,54; 2,53]	1,99 (0,25)	[1,53; 2,54]	2,3,5,6,9,1
RR[5] Tâmega	1842	6410	0,29	0,29 (0,04)	[0,23 ; 0,37]	0,29 (0,04)	[0,22; 0,37]	0,29 (0,04)	[0,22; 0,37]	3,4,6,7,8,13
RR[6] Entre Douro e Vouga	NA	3855	-	1,63 (1,70)	[0,25 ; 5,61]	1,74 (1,86)	[0,26; 6,16]	1,64 (1,72)	[0,23; 5,86]	4,5,7,13,9
RR[7] Douro	3452	3438	1,00	1,02 (0,13)	[0,79 ; 1,29]	1,01 (0,13)	[0,78; 1,29]	1,01 (0,13)	[0,78; 1,29]	6,8,5,13,17
RR[8] Alto Trás-os-Montes	NA	4095	-	0,81 (0,94)	[0,11 ; 2,95]	0,71 (0,84)	[0,09; 2,64]	0,90 (1,13)	[0,11; 3,35]	1,2,3,5,7
RR[9] Baixo Vouga	9643	5980	1,61	1,64 (0,20)	[1,27 ; 2,07]	1,63 (0,21)	[1,25; 2,07]	1,62 (0,21)	[1,25; 2,07]	4,6,13,10,12,
RR[10] Baixo Mondego	1951	5619	0,35	0,35 (0,04)	[0,27 ; 0,45]	0,35 (0,05)	[0,27; 0,45]	0,35 (0,05)	[0,27; 0,45]	9,13,12,11
RR[11] Pinhal Litoral	1328	4145	0,32	0,33 (0,04)	[0,25 ; 0,41]	0,32 (0,04)	[0,25; 0,41]	0,32 (0,04)	[0,25; 0,41]	10,12,22,23,19
RR[12] Pinhal Interior Norte	2290	2576	0,89	0,90 (0,11)	[0,70 ; 1,15]	0,90 (0,11)	[0,69; 1,14]	0,90 (0,12)	[0,69; 1,15]	10,9,13,15,16,14,22,11,18
RR[13] Dão-Lafões	11906	4900	2,43	2,47 (0,31)	[1,92 ; 3,12]	2,45 (0,31)	[1,89; 3,12]	2,45 (0,31)	[1,89; 3,12]	10,9,6,5,7,17,15,12
RR[14] Pinhal Interior Sul	1344	910	1,48	1,50 (0,19)	[1,16 ; 1,91]	1,45 (0,19)	[1,14; 1,9]	1,49 (0,19)	[1,14; 1,91]	22,12,16,18,25
RR[15] Serra da Estrela	6128	882	6,95	7,05 (0,88)	[5,47 ; 8,94]	7,00 (0,89)	[5,4; 8,91]	6,99 (0,90)	[5,39; 8,92]	13,17,16,12
RR[16] Beira Interior Norte	NA	2120	-	2,83 (3,17)	[0,38 ; 10,30]	2,82 (3,24)	[0,36; 10,41]	3,26 (3,84)	[0,42; 12,06]	18,14,12,15,17
RR[17] Beira Interior Sul	NA	1531	-	3,10 (3,33)	[0,43 ; 11,11]	2,91 (3,81)	[0,38; 10,66]	3,15 (3,77)	[0,41; 11,68]	7,13,15,16,18
RR[18] Cova da Beira	NA	1647	-	2,07 (2,18)	[0,29 ; 7,35]	2,22 (2,49)	[0,3; 8,02]	2,08 (2,38)	[0,27; 7,69]	25,14,17,16,12
RR[19] Oeste	4221	5769	0,73	0,74 (0,09)	[0,58 ; 0,94]	0,74 (0,09)	[0,57; 0,94]	0,74 (0,09)	[0,57; 0,94]	11,23,20
RR[20] Médio Tejo	1950	4038	0,48	0,49 (0,06)	[0,38 ; 0,62]	0,49 (0,06)	[0,37; 0,62]	0,49 (0,06)	[0,37; 0,62]	19,23,21
RR[21] Grande Lisboa	35207	30615	1,15	1,17 (0,14)	[0,91 ; 1,48]	1,16 (0,15)	[0,89; 1,47]	1,16 (0,15)	[0,89; 1,48]	20,23,26,24
RR[22] Península de Setúbal	NA	11552	-	1,27 (1,28)	[0,19 ; 4,39]	1,48 (1,60)	[0,22; 5,23]	1,22 (1,33)	[0,17; 4,32]	25,23,11,12,14
RR[23] Alentejo Litoral	2302	1802	1,28	1,30 (0,16)	[1,00 ; 1,64]	1,23 (0,16)	[0,99; 1,64]	1,29 (0,17)	[0,99; 1,64]	22,11,19,20,21,26,25
RR[24] Alto Alentejo	720	2320	0,31	0,32 (0,04)	[0,24 ; 0,40]	0,31 (0,04)	[0,24; 0,4]	0,31 (0,04)	[0,24; 0,4]	21,26,27,28
RR[25] Alentejo Central	3831	3182	1,20	1,22 (0,15)	[0,95 ; 1,55]	1,21 (0,15)	[0,94; 1,55]	1,21 (0,16)	[0,93; 1,55]	18,14,22,23,26
RR[26] Baixo Alentejo	600	2302	0,26	0,27 (0,03)	[0,20 ; 0,34]	0,26 (0,03)	[0,2; 0,34]	0,26 (0,04)	[0,2; 0,34]	25,23,21,24,27
RR[27] Lezíria do Tejo	3885	4280	0,91	0,92 (0,11)	[0,71 ; 1,17]	0,92 (0,12)	[0,71; 1,17]	0,91 (0,12)	[0,7; 1,17]	26,24,28
RR[28] Algarve	2886	6840	0,42	0,43 (0,05)	[0,33 ; 0,54]	0,43 (0,05)	[0,33; 0,54]	0,43 (0,05)	[0,33; 0,54]	24,27
RR[29] RAA	2091	2697	0,78	0,90 (0,09)	[0,73 ; 1,09]	1,06 (0,10)	[0,87; 1,27]	1,06 (0,1)	[0,88; 1,26]	-
RR[30] RAM	3792	2867	1,32	0,90 (0,09)	[0,73 ; 1,09]	1,06 (0,10)	[0,87; 1,27]	1,08 (0,1)	[0,9; 1,28]	-

Tabela 6.11: Estimativas de Risco Relativo por NUTS III

Os seguintes mapas apresentam a informação de risco relativo por NUTS III, estimado para o modelo sem covariáveis (6.9), modelo incluindo a idade como covariável (6.10) e modelo considerando o IMC (6.11).

Os mapas de risco relativo apresentados são muito sobreponíveis entre si, evidenciando maior risco na região centro do país devido às regiões do Pinhal Interior Norte e Sul, Beira Interior Norte e Sul e Serra da Estrela. A sul do país destaca-se a região do Algarve com um risco relativo mais elevado, em todos os modelos. É na região de Alto de Trás-os-Montes onde os mapas apresentam uma ligeira variação.

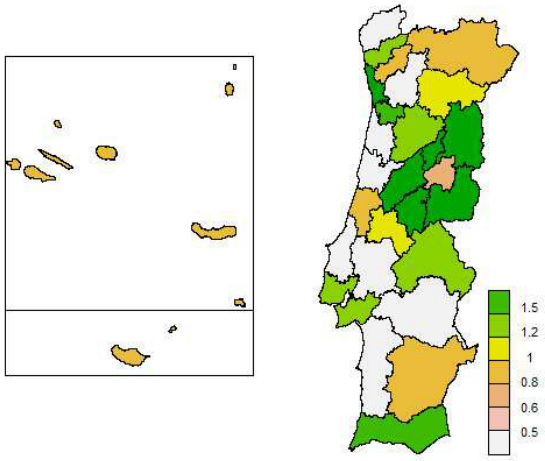


Figura 6.9: Risco Relativo por NUTS III - modelo sem covariáveis

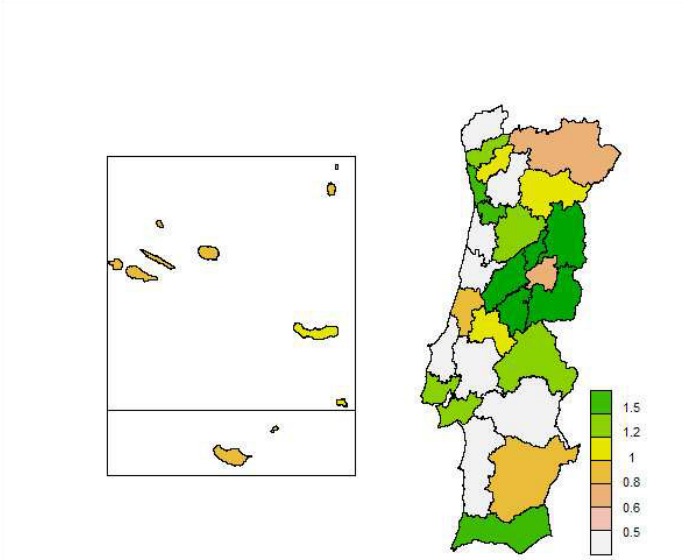


Figura 6.10: Risco Relativo por NUTS III - modelo incluindo a Idade

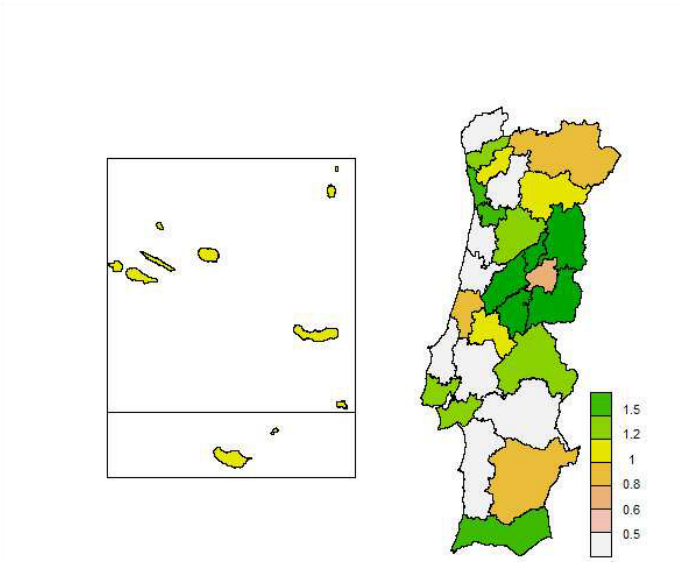


Figura 6.11: Risco Relativo por NUTS III - modelo incluindo o IMC

6.3 Prevalência de FA por NUTS III

Os resultados de prevalência de FA por NUTS III, obtidos através do modelo de BYM, apresentam variações significativas entre regiões. A região do Pinhal Interior Sul destaca-se como sendo a região com maior prevalência de FA, aproximadamente 23%, seguida da região da Serra da Estrela com cerca de 10% de prevalência. As regiões onde o valor de prevalência estimado é menor situam-se no Baixo Mondego, Alentejo Central e Tâmega. Também os mapas para a prevalência de FA apresentam-se muito semelhantes entre os vários modelos estudados, apresentado, no entanto, algumas alterações para a região de Alto de Trás-os-Montes, Grande Lisboa, RAA e RAM.

		Modelo sem Covariáveis		Modelo sem Covariáveis: Idade		Modelo sem Covariáveis: IMC		Regiões Adjacentes
NUTS III	Prevalência [i]	Média (DP)	IC 95%	Média (DP)	IC 95%	Média (DP)	IC 95%	
TP[1] Minho-Lima	0,036	0,014 (0,454)	[0,013; 0,015]	0,014 (0,453)	[0,013; 0,015]	0,014 (0,452)	[0,013; 0,015]	2,8,3,4
TP[2] Cávado	NA	0,031 (0,044)	[0,003; 0,125]	0,034 (0,051)	[0,003; 0,141]	0,032 (0,047)	[0,003; 0,138]	1,3,8,4
TP[3] Ave	NA	0,023 (0,027)	[0,003; 0,085]	0,025 (0,03)	[0,003; 0,095]	0,025 (0,028)	[0,003; 0,092]	2,4,5,8,1
TP[4] Grande Porto	0,023	0,053 (0,007)	[0,041; 0,067]	0,053 (0,007)	[0,041; 0,067]	0,053 (0,007)	[0,041; 0,067]	2,3,5,6,9,1
TP[5] Tâmega	0,013	0,007 (0,938)	[0,006; 0,009]	0,007 (0,952)	[0,006; 0,009]	0,007 (0,959)	[0,006; 0,009]	3,4,6,7,8,13
TP[6] Entre Douro e Vouga	NA	0,043 (0,044)	[0,006; 0,147]	0,046 (0,049)	[0,007; 0,161]	0,043 (0,05)	[0,006; 0,154]	4,5,7,13,9
TP[7] Douro	0,048	0,032 (0,004)	[0,025; 0,041]	0,032 (0,004)	[0,025; 0,041]	0,032 (0,004)	[0,024; 0,041]	6,8,5,13,17
TP[8] Alto Trás-os-Montes	NA	0,028 (0,032)	[0,004; 0,101]	0,024 (0,029)	[0,003; 0,091]	0,031 (0,033)	[0,004; 0,116]	1,2,3,5,7
TP[9] Baixo Vouga	0,058	0,047 (0,006)	[0,036; 0,059]	0,047 (0,006)	[0,036; 0,059]	0,047 (0,006)	[0,036; 0,059]	4,6,13,10,12,
TP[10] Baixo Mondego	0,023	0,011 (0,001)	[0,008; 0,014]	0,011 (0,001)	[0,008; 0,014]	0,011 (0,001)	[0,008; 0,014]	9,13,12,11
TP[11] Pinhal Litoral	0,03	0,010 (0,001)	[0,007; 0,012]	0,01 (0,001)	[0,007; 0,012]	0,01 (0,001)	[0,007; 0,012]	10,12,22,23,19
TP[12] Pinhal Interior Norte	0,02	0,030 (0,004)	[0,023; 0,039]	0,03 (0,004)	[0,023; 0,039]	0,03 (0,004)	[0,023; 0,038]	10,9,13,15,16,14,22,11,18
TP[13] Dão-Lafões	0,033	0,080 (0,001)	[0,062; 0,101]	0,079 (0,01)	[0,061; 0,101]	0,079 (0,01)	[0,061; 0,1]	10,9,6,5,7,17,15,12
TP[14] Pinhal Interior Sul	0,012	0,057 (0,007)	[0,044; 0,073]	0,057 (0,007)	[0,043; 0,073]	0,057 (0,008)	[0,043; 0,073]	22,12,16,18,25
TP[15] Serra da Estrela	0,1	0,234 (0,029)	[0,181; 0,297]	0,232 (0,03)	[0,179; 0,296]	0,232 (0,03)	[0,179; 0,295]	13,17,16,12
TP[16] Beira Interior Norte	NA	0,097 (0,109)	[0,013; 0,352]	0,097 (0,112)	[0,013; 0,36]	0,114 (0,136)	[0,015; 0,414]	18,14,12,15,17
TP[17] Beira Interior Sul	NA	0,111 (0,121)	[0,015; 0,4]	0,104 (0,136)	[0,014; 0,38]	0,111 (0,124)	[0,014; 0,424]	7,13,15,16,18
TP[18] Cova da Beira	Na	0,067 (0,071)	[0,009; 0,242]	0,072 (0,081)	[0,01; 0,261]	0,069 (0,077)	[0,009; 0,262]	25,14,17,16,12
TP[19] Oeste	0,033	0,022 (0,003)	[0,017; 0,028]	0,022 (0,003)	[0,017; 0,028]	0,022 (0,003)	[0,017; 0,028]	11,23,20
TP[20] Médio Tejo	0,03	0,016 (0,002)	[0,012; 0,02]	0,016 (0,002)	[0,012; 0,02]	0,016 (0,002)	[0,012; 0,02]	19,23,21
TP[21] Grande Lisboa	0,022	0,034 (0,004)	[0,026; 0,043]	0,034 (0,004)	[0,026; 0,043]	0,034 (0,004)	[0,026; 0,043]	20,23,26,24
TP[22] Península de Setúbal	NA	0,036 (0,04)	[0,005; 0,124]	0,042 (0,045)	[0,006; 0,147]	0,034 (0,034)	[0,005; 0,119]	25,23,11,12,14
TP[23] Alentejo Litoral	0,035	0,043 (0,005)	[0,033; 0,055]	0,043 (0,006)	[0,033; 0,055]	0,043 (0,006)	[0,033; 0,055]	22,11,19,20,21,26,25
TP[24] Alto Alentejo	0,045	0,011 (0,001)	[0,008; 0,014]	0,011 (0,002)	[0,008; 0,014]	0,011 (0,002)	[0,008; 0,014]	21,26,27,28
TP[25] Alentejo Central	0,022	0,041 (0,005)	[0,032; 0,053]	0,041 (0,005)	[0,032; 0,052]	0,041 (0,005)	[0,032; 0,052]	18,14,22,23,26
TP[26] Baixo Alentejo	0,023	0,009 (0,001)	[0,007; 0,011]	0,009 (0,001)	[0,007; 0,011]	0,009 (0,001)	[0,007; 0,011]	25,23,21,24,27
TP[27] Lezíria do Tejo	0,04	0,029 (0,004)	[0,023; 0,037]	0,029 (0,004)	[0,022; 0,037]	0,029 (0,004)	[0,022; 0,037]	26,24,28
TP[28] Algarve	0,013	0,013 (0,002)	[0,01; 0,016]	0,013 (0,002)	[0,01; 0,016]	0,013 (0,002)	[0,01; 0,016]	24,27
TP[29] RAA	0,02	0,023 (0,002)	[0,019; 0,028]	0,027 (0,003)	[0,022; 0,033]	0,027 (0,003)	[0,023; 0,032]	-
TP[30] RAM	0,034	0,023 (0,002)	[0,019; 0,028]	0,028 (0,003)	[0,023; 0,033]	0,028 (0,003)	[0,023; 0,033]	-

Tabela 6.12: Prevalência de FA por NUTS III

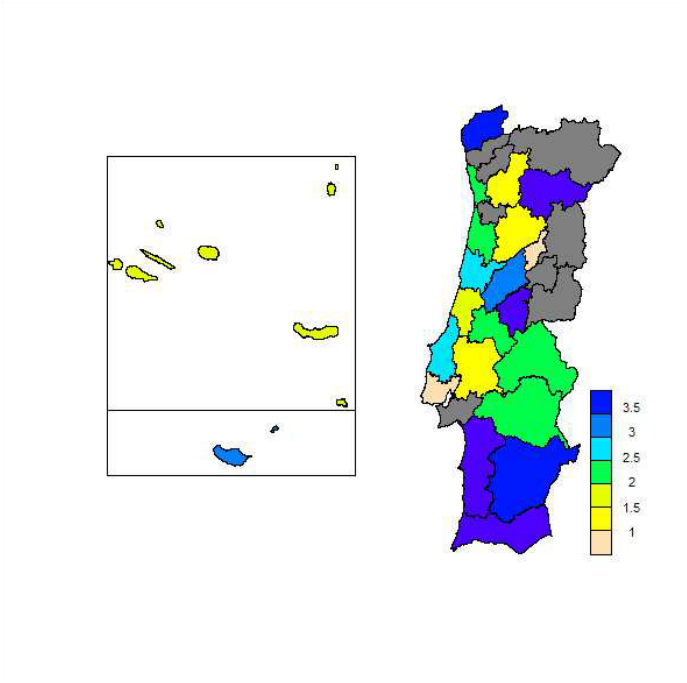


Figura 6.12: Taxa de Prevalência de FA por NUTS III (%) - modelo frequencista

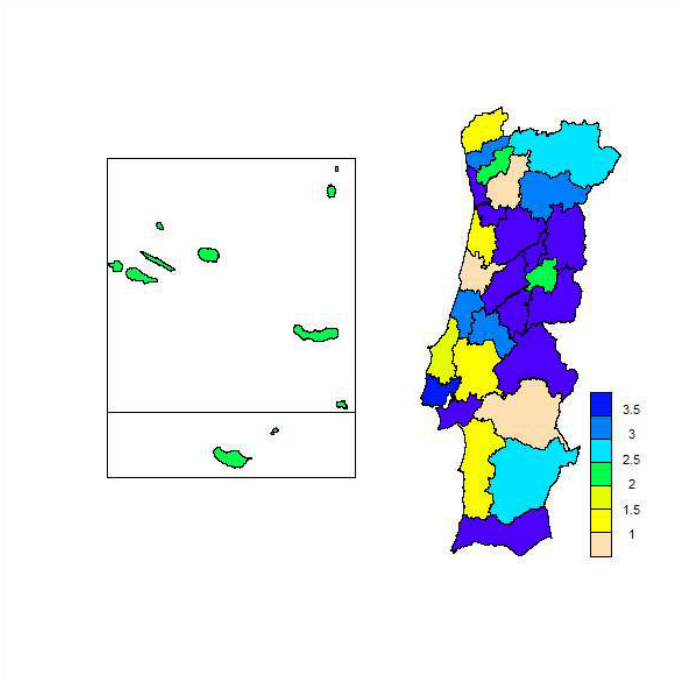


Figura 6.13: Taxa de Prevalência de FA por NUTS III (%) - modelo sem covariáveis

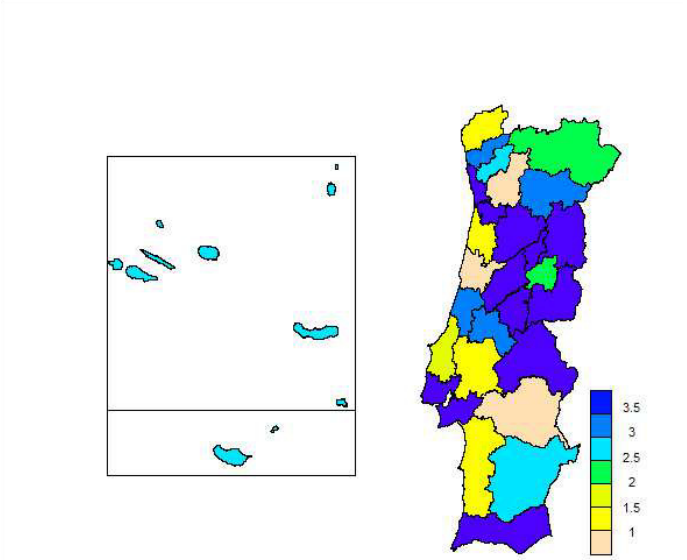


Figura 6.14: Taxa de Prevalência de FA por NUTS III (%) - modelo incluindo a Idade

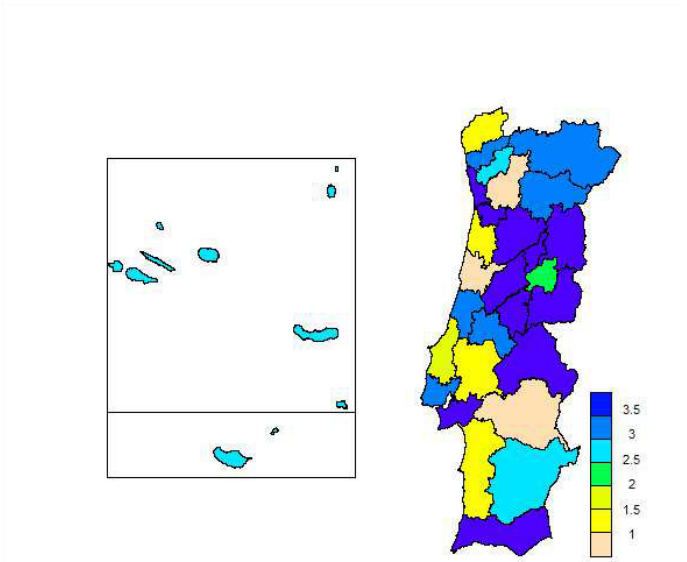


Figura 6.15: Taxa de Prevalência de FA por NUTS III (%) - modelo incluindo o IMC

Capítulo 7

Discussão dos Resultados

A Fibrilhação Auricular pelo impacto que tem na sociedade devido às comorbilidade associadas e elevada mortalidade, tem vindo a ganhar relevância clínica e epidemiológica. Em 2010 foi realizado um estudo, de âmbito nacional, para estimar o valor de prevalência da doença na população portuguesa com idade igual ou superior a 40 anos. Os dados de prevalência publicados indicaram uma estimativa de 2,5%, com um aumento crescente ao nível das classes etárias, em conformidade com os vários estudos publicados. Foram observadas variações regionais, sobre as quais não foram feitas muitas considerações. As limitações da análise frequencista e a sobre dispersão dos dados, consequência de um número relativamente pequeno de observações com FA, levou a que as variações regionais não fossem explicadas.

A detecção de erros no desenho de estudo levou, a que numa primeiro passo, todos os resultados do estudo FAMA fossem revistos em função de um processo de pós-estratificação dos dados, realizado no âmbito deste estudo. A estimativas encontradas para o total de casos de FA nesta população foi de 138.427, muito superior ao publicado e que estimava um

valor de 121.825 casos. A justificação para esta discrepância de valores na estimativa total de casos de FA está relacionada com os pesos de inclusão de cada indivíduo na amostra, que se encontravam desajustados e como tal não forneceram estimativas próximas da realidade. Este número de casos estimados pode ainda ser mais elevado do que o estimado, porque por razões clínicas casos de FA paroxística podem não ter sido identificados e documentados. No entanto, este argumento, já havia sido referido pelos autores do estudo.

A taxa de prevalência estimada para FA foi de 2,52%, próxima da já publicada. Entre homens e mulheres a taxa de prevalência é diferente mas não de forma significativa, em linha com os resultados já descritos pelo estudo, muito embora as taxas apresentadas fossem idênticas entre homens e mulheres. Também na estimação da prevalência de FA por classes etárias observam-se taxas próximas do já publicado, confirmando-se o aumento significativo da prevalência de FA com o aumento da idade.

É na estimação dos valores de prevalência de FA por região onde os resultados estimados são contraditórios com os resultados apresentados. Foram identificadas diferenças estatisticamente significativas nos valores observados de FA por NUTS II, mais elevados na região Centro, Alentejo e Região Autónoma da Madeira, mais baixas sobretudo na região do Algarve e na Região Autónoma dos Açores, mas estas variações não foram detectadas nos resultados publicados. Uma das questões levantadas na interpretação destes resultados é a elevada variabilidade e sobre dispersão dos dados, decorrente da abordagem frequencista. Esta é a justificação para a utilização de modelação Bayesiana.

Através da utilização de modelos hierárquicos Bayesianos foi possível ultrapassar as limitações da abordagem frequencista e identificar a existência de uma estrutura espacial nos dados, que determina a existência de variações a nível regional. Os dados foram modelados considerando o modelo construído por Besag, York e Mollié, e, desde logo são assumidas as vantagens na sua utilização. A estimação de valores para regiões para as quais não existe informação disponível, quer de valores de risco relativo quer de valores de prevalência da doença, é claramente uma vantagem comparativamente à utilização da abordagem frequencista, apenas conseguida porque este modelo considera uma componente de estrutura espacial dos dados, e que através dos valores de regiões vizinhas ou contíguas a regiões vizinhas estima os valores em falta para essas regiões. No entanto, a análise destas estimativas deve ser sempre feita com algum cuidado, uma vez que apresentam maior dispersão ou variabilidade comparativamente às restantes regiões para as quais havia informação disponível, tal como identificado neste estudo.

Para além da estimação de valores para regiões sem informação, a utilização de métodos Bayesianos no mapeamento de doenças permite obter estimativas de parâmetros epidemiológicos, quer seja risco relativo quer seja taxas de prevalência ou incidência, com algum grau de alisamento, sendo possível identificar padrões na variação dos resultados. Os mapas apresentados com base nas estimativas de risco relativo e prevalência de FA demonstram este facto, quando comparados com os mapas construídos com base nas estimativas do método clássico. Observou-se um cluster na região Centro, que justifica uma prevalência mais elevada de FA nesta região, que inclui as regiões do Pinhal Interior Norte e Sul, Beira Interior Norte e Sul e Serra da Estrela. Embora menos evidente, nas regiões de zonas litorais os valores de risco apresentam-se mais baixos, com excepção do Algarve. Este padrão não é sobreponível

à análise dos mapas obtidos pelo método clássico.

A possibilidade de modelação dos dados considerando informação externa e que permite obter um melhor ajustamento, é também um factor diferenciador na utilização deste tipo de modelos. No entanto, ao contrário do expectável, a utilização de covariáveis neste modelo não revelou ser uma mais valia, uma vez que as estimativas obtidas apresentaram pouca variação nos resultados. O valor de *deviance* dos modelos estimados revela este facto. A escolha das variáveis a incluir no modelo baseou-se na sua identificação enquanto factor de risco. Contudo, o facto de ser factor de risco para a doença não implica que ao nível do ajustamento dos dados por região tenha o mesmo impacto. Quando analisado os coeficientes de ponderação das variáveis incluídas no modelo, percentagem de indivíduos com idade igual ou superior a 65 anos e valor médio de IMC, verificou-se que são relativamente baixos, não tendo uma grande importância na determinação do risco ou prevalência de FA. A interpretação dos coeficientes de ponderação também são alvo de alguma discussão, sobretudo ao nível da idade, por se apresentarem contraditórios com o esperado. O sinal negativo associado ao coeficiente indica que um aumento de uma unidade percentual na variável, ou seja, um aumento de 1% na percentagem de pessoas com 65 ou mais anos de idade na região faz decrescer o valor de risco associado. O esperado, considerando que o valor de FA aumenta significativamente com a idade, seria que este valor contribuísse de forma positiva para aumentar o risco relativo determinado. A interpretação deste parâmetro não é clara no verdadeiro impacto da idade no risco da doença, mas análises adicionais realizadas com a variável idade mantiveram sempre um valor negativo no sinal do sinal do ponderador. Mais ainda, a análise de sensibilidade realizada não foi contraditória com os resultados apresentados pela análise principal. No que

respeita ao valor de IMC, a interpretação é mais consonante, uma vez que o sinal associado ao parâmetro é positivo, indicando que o incremente do valor médio de IMC contribui para aumentar o valor de risco relativo da doença.

Capítulo 8

Conclusões

Os resultados deste estudo demonstram que o uso de modelos Bayesianos são uma melhor opção para a estimação de valores de prevalência de uma doença, risco relativo ou mortalidade quando o número de casos observados é pequeno, permitindo estimar informação para regiões sobre as quais tal não estava disponível, somente possível devido à utilização de um modelo que capta a estrutura espacial dos dados.

Neste estudo, a utilização de covariáveis no modelo não trouxe um valor acrescentado na precisão dos estimadores de risco relativo e prevalência de Fibrilhação Auricular, mas a possibilidade de modelar a componente espacial dos dados permitiu, não só estimar valores para regiões onde essa informação não estava disponível, mas também identificar uma variação regional nas estimativas encontradas

Assim, do estudo FAMA conclui-se que a Fibrilhação Auricular apresenta variações regionais, a nível de NUTS III, que não devem ser desvalorizadas na determinação de políticas de saúde

pública para controle da doença.

Capítulo 9

Limitações e Recomendações para trabalhos futuros

A utilização de modelos Bayesianos nos dados do estudo FAMA permitiu colmatar a limitação da abordagem frequencista na estimação de valores de prevalência da doença num cenário onde o número de observações é pequeno. As vantagens associadas à modelação Bayesiana são reconhecidas, e o modelo proposto por Besag, York e Mollié, numa perspectiva totalmente Bayesiana, conferiu a esta abordagem uma mais valia pela possibilidade de introduzir uma componente espacial no modelo estruturado. Em 2004, Richardson e colegas (2004) publicaram um artigo sobre a interpretação das estimativas *a posteriori* do risco relativo no mapeamento de doenças, referindo que uma das limitações do modelo proposto por BYM é a baixa sensibilidade para a detecção de risco relativo emergente, e consequentemente a real magnitude do risco encontra-se subvalorizada nestes modelos. Os autores do artigo questionam também a escolha das priors utilizadas para o alisamento dos dados na produção das estimativas de risco por região.

O modelo proposto por Besag, York e Mollié tem vindo a ser actualizado por outro autores (Gelfand & Vounatsou, 2003; Knorr-Held & Best, 2001) para melhor a sua performance quanto às estimativas de risco relativo *a posteriori*. Estes novos modelos consideram que, para além da estrutura espacial dos dados, é necessário incluir uma componente de informação relativa a outras patologias associadas ou factores de risco, com uma etiologia semelhante à da doença em estudo, e com um padrão de dispersão geográfico semelhante, para que a modelação da informação seja melhor ajustada e melhor explique a variação dos dados. Assim, para cada doença, o risco relativo numa dada região i é modelado como o produto de uma componente partilhada da doença θ_i e a própria doença ϕ_{di}:

$$y_{1_i} \sim P(e_{1_i}\theta_i\delta_{1_i}^{\phi})$$

$$y_{2_i} \sim P(e_{2_i}\theta_i\delta_{2_i}^{\phi})$$

Nos dados do estudo FAMA foram avaliadas outras comorbilidades tais como a HTA, clinicamente relacionadas com a FA e identificadas como factor de risco, mas que através do modelo proposto por BYM não foi possível modelar por falta de convergência dos parâmetros, sendo uma limitação reconhecida. Como proposta de trabalho sugere-se a revisão do modelo aqui apresentado em função da proposta de Gelfand e Vounatsou (2003) ou Knorr-Held e Best (2001) para a estimação do valor de risco relativo de FA, dada as variações encontradas ao nível da NUTS III.

Tal como referido anteriormente, a escolha das priors poderá ser revista. Esta é uma das limitações da análise Bayesiana, uma vez que se encontra dependente da informação dispo-

nível e do conhecimento que se tem sobre o fenómeno em estudo. A escolha subjectiva desta informação é viés nos modelos determinados.

Anexo 1

Valores marginais utilizados na pós estratificação

NUTS_II	SEXO	PS_EE	Freq
NORTE	Feminino	40-44	145945
NORTE	Feminino	45-49	149051
NORTE	Feminino	50-55	134814
NORTE	Feminino	55-59	123127
NORTE	Feminino	60-64	107864
NORTE	Feminino	65-69	92921
NORTE	Feminino	70-74	85758
NORTE	Feminino	75-79	76840
NORTE	Feminino	80-84	55892
NORTE	Feminino	85+	41645
NORTE	Masculino	40-44	140125
NORTE	Masculino	45-49	141563
NORTE	Masculino	50-55	127237
NORTE	Masculino	55-59	112532
NORTE	Masculino	60-64	96687
NORTE	Masculino	65-69	76572
NORTE	Masculino	70-74	66516
NORTE	Masculino	75-79	53865
NORTE	Masculino	80-84	32826
NORTE	Masculino	85+	19963

NUTS_II	SEXO	PS_EE	Freq
CENTRO	Feminino	40-44	62789
CENTRO	Feminino	45-49	65100
CENTRO	Feminino	50-55	61801
CENTRO	Feminino	55-59	57615
CENTRO	Feminino	60-64	54395
CENTRO	Feminino	65-69	51795
CENTRO	Feminino	70-74	51503
CENTRO	Feminino	75-79	48763
CENTRO	Feminino	80-84	36894
CENTRO	Feminino	85+	29690
CENTRO	Masculino	40-44	61832
CENTRO	Masculino	45-49	62860
CENTRO	Masculino	50-55	59603
CENTRO	Masculino	55-59	54410
CENTRO	Masculino	60-64	48631
CENTRO	Masculino	65-69	42470
CENTRO	Masculino	70-74	40748
CENTRO	Masculino	75-79	35480
CENTRO	Masculino	80-84	23039
CENTRO	Masculino	85+	15127

NUTS_II	SEXO	PS_EE	Freq
LVT	Feminino	40-44	134790
LVT	Feminino	45-49	132178
LVT	Feminino	50-55	126036
LVT	Feminino	55-59	123899
LVT	Feminino	60-64	122808
LVT	Feminino	65-69	108663
LVT	Feminino	70-74	96010
LVT	Feminino	75-79	85659
LVT	Feminino	80-84	62091
LVT	Feminino	85+	46327
LVT	Masculino	40-44	134491
LVT	Masculino	45-49	125688
LVT	Masculino	50-55	116382
LVT	Masculino	55-59	109232
LVT	Masculino	60-64	106139
LVT	Masculino	65-69	92096
LVT	Masculino	70-74	77400
LVT	Masculino	75-79	60458
LVT	Masculino	80-84	36066
LVT	Masculino	85+	20976

NUTS_II	SEXO	PS_EE	Freq
ALENTEJO	Feminino	40-44	16070
ALENTEJO	Feminino	45-49	17139
ALENTEJO	Feminino	50-55	16923
ALENTEJO	Feminino	55-59	15636
ALENTEJO	Feminino	60-64	15026
ALENTEJO	Feminino	65-69	14430
ALENTEJO	Feminino	70-74	16477
ALENTEJO	Feminino	75-79	16416
ALENTEJO	Feminino	80+	21604
ALENTEJO	Masculino	40-44	17161
ALENTEJO	Masculino	45-49	18021
ALENTEJO	Masculino	50-55	17566
ALENTEJO	Masculino	55-59	15978
ALENTEJO	Masculino	60-64	13789
ALENTEJO	Masculino	65-69	11855
ALENTEJO	Masculino	70-74	12711
ALENTEJO	Masculino	75-79	12379
ALENTEJO	Masculino	80+	14017

NUTS_II	SEXO	PS_EE	Freq
ALGARVE	Feminino	40-44	15308
ALGARVE	Feminino	45-49	15413
ALGARVE	Feminino	50-55	14612
ALGARVE	Feminino	55-59	13843
ALGARVE	Feminino	60-64	12745
ALGARVE	Feminino	65-69	11589
ALGARVE	Feminino	70-74	11049
ALGARVE	Feminino	75-79	10510
ALGARVE	Feminino	80+	13616
ALGARVE	Masculino	40-44	16713
ALGARVE	Masculino	45-49	15990
ALGARVE	Masculino	50-55	15116
ALGARVE	Masculino	55-59	13783
ALGARVE	Masculino	60-64	12457
ALGARVE	Masculino	65-69	10783
ALGARVE	Masculino	70-74	9851
ALGARVE	Masculino	75+	17065

NUTS_II	SEXO	PS_EE	Freq
RAM	Feminino	40-44	9977
RAM	Feminino	45-49	10062
RAM	Feminino	50-55	8321
RAM	Feminino	55-59	7292
RAM	Feminino	60-64	6146
RAM	Feminino	65-69	5534
RAM	Feminino	70+	15688
RAM	Masculino	40-44	9376
RAM	Masculino	45-49	8785
RAM	Masculino	50-59	13182
RAM	Masculino	60+	15805

NUTS_II	SEXO	PS_EE	Freq
RAA	Feminino	40-49	17268
RAA	Feminino	50-59	14248
RAA	Feminino	60-69	10507
RAA	Feminino	70+	13390
RAA	Masculino	40-49	17765
RAA	Masculino	50-59	14652
RAA	Masculino	60+	17555

Anexo 2

Análise Bayesiana - Modelo I (sem covariáveis)

```
model
{
for (i in 1 : N) {
O[i] ∼ dpois(mu[i])
log(mu[i]) <- log(E[i]) + alpha0 + b[i]
RR[i] <- exp(alpha0 + b[i])
O1[i]  dpois(mu[i])
tp[i]<-O1[i]/PR[i]
}

b[1:N]   car.normal(adj[], weights[], num[], tau)
for(k in 1:sumNumNeigh) {
weights[k] <- 1
}

alpha0 ∼ dflat()
tau ∼ dgamma(0.5, 0.05)
sigma <- sqrt(1 / tau)
}
```

Análise Bayesiana - Modelo I (dados)

list(N = 30,

O = c(1826, NA, NA, 35233, 1842, NA, 3452, NA, 9643, 1951, 1328,2290, 11906, 1344, 6128, NA, NA, NA, 4221, 1950, 35207,

NA,2302,720, 3831, 600, 3885, 2886, 2091, 3792),

E = c(4280.316824, 4901.15734, 6372.499508, 17838.34636, 6410.199816,3854.729947, 3437.686045, 4094.987028, 5979.801876,

5619.080017,4145.312634, 2575.571111, 4899.673362, 909.9561532, 881.759415, 2120.157603, 1530.881776, 1646.723415,

5768.649325,4037.953121,30614.5844, 11551.67259, 1801.959501, 2320.303559, 3181.540394,2302.47843, 4279.516298,

6840.330412, 2696.549893, 2867.133147),

PR=c(132830, 192261, 254748, 671340, 254636, 147624, 109093, 119211, 208575, 181582, 140706, 76576, 151787, 23927,

26566, 61387, 42842, 50597, 191401, 125971, 1054768, 409811, 54021, 66149, 93974, 69054, 135438, 230443, 105385, 110168),

num = c(4,4,5,6,6,5,5,5,5,4,5,9,8,5,4,5,5,5,3,3,4,5,7,4,5,5,3,2,0,0),

adj = c(2,8,3,4,

1,3,8,4,

2,4,5,8,1,

2,3,5,6,9,1,

3,4,6,7,8,13,

4,5,7,13,9,

6,8,5,13,17,

1,2,3,5,7,

4,6,13,10,12,

9,13,12,11,

10,12,22,23,19,

10,9,13,15,16,14,22,11,18,

10,9,6,5,7,17,15,12,

22,12,16,18,25,

13,17,16,12,

18,14,12,15,17,

7,13,15,16,18,

25,14,17,16,12,

11,23,20,

19,23,21,

20,23,26,24,

25,23,11,12,14,

22,11,19,20,21,26,25,

21,26,27,28,

18,14,22,23,26,

25,23,21,24,27,

26,24,28,

24,27),

sumNumNeigh = 136)

Análise Bayesiana - Modelo I (inits)

```
list(tau = 1, alpha0 = 0,

b=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0))
```

Análise Bayesiana - Modelo II (com covariáveis: % indivíduos com >=65 anos)

```
model
{
for (i in 1 : N) {
O[i] ∼ dpois(mu[i])
log(mu[i]) <- log(E[i]) + alpha0 + alpha1 * idade[i] + b[i]
RR[i] <- exp(alpha0 + alpha1 * idade[i] + b[i])
O1[i]  dpois(mu[i])
tp[i]<-O1[i]/PR[i]
}
b[1:N]   car.normal(adj[], weights[], num[], tau)
for(k in 1:sumNumNeigh) {
weights[k] <- 1
}

alpha0 ∼ dflat()
alpha1 ∼ dnorm(0.0, 1.0E-5)
tau ∼ dgamma(0.5, 0.05)
sigma <- sqrt(1 / tau)
}
```

Análise Bayesiana - Modelo II (dados)

```
list(N = 30, O = c(1826, NA, NA, 35233, 1842, NA, 3452, NA, 9643, 1951, 1328,2290, 11906, 1344, 6128, NA, NA, NA, 4221, 1950,
35207, NA, 2302,720, 3831, 600, 3885, 2886, 2091, 3792),
E = c( 4280.316824, 4901.15734, 6372.499508, 17838.34636, 6410.199816,3854.729947, 3437.686045, 4094.987028, 5979.801876,
5619.080017,4145.312634, 2575.571111, 4899.673362, 909.9561532, 881.759415,2120.157603, 1530.881776, 1646.723415,
5768.649325, 4037.953121,30614.5844, 11551.67259, 1801.959501, 2320.303559, 3181.540394,2302.47843, 4279.516298,
6840.330412,2696.549893, 2867.133147),
idade=c(0.374950253, 0.707581773, -0.404396217, -0.884026222, -0.83133438, 0.741178314, -0.241720858,1.381620143,
-0.410175312,0.388277203, -0.922473415, 0.626289638, -0.418983047, 0.837938994,0.61228346, 0.890038074, 0.506002529,
-3.928740855, 0.136347032, 0.234902113, -0.486875105,0.375142105, 0.098122467, 1.121793064, 0.351968935, 0.769313338,
0.61853095, -0.260844375,-1.159335242, -0.823375357),
PR=c(132830, 192261, 254748, 671340, 254636, 147624, 109093, 119211, 208575, 181582, 140706, 76576, 151787, 23927,
26566, 61387, 42842, 50597, 191401, 125971, 1054768, 409811, 54021, 66149, 93974, 69054, 135438, 230443, 105385, 110168),
num = c(4,4,5,6,6,5,5,5,5,4,5,9,8,5,4,5,5,5,3,3,4,5,7,4,5,5,3,2,0,0),
adj = c(2,8,3,4,
1,3,8,4,
2,4,5,8,1,
2,3,5,6,9,1,
```

3,4,6,7,8,13,

4,5,7,13,9,

6,8,5,13,17,

1,2,3,5,7,

4,6,13,10,12,

9,13,12,11,

10,12,22,23,19,

10,9,13,15,16,14,22,11,18,

10,9,6,5,7,17,15,12,

22,12,16,18,25,

13,17,16,12,

18,14,12,15,17,

7,13,15,16,18,

25,14,17,16,12,

11,23,20,

19,23,21,

20,23,26,24,

25,23,11,12,14,

22,11,19,20,21,26,25,

21,26,27,28,

18,14,22,23,26,

25,23,21,24,27,

26,24,28,

24,27),

sumNumNeigh = 136)

Análise Bayesiana - Modelo II (inits)

list(tau = 1, alpha0 = 0, alpha1 = 0,

b=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0))

list(tau = 1, alpha0 = 0, alpha1 = 0,

b=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0))

Análise Bayesiana - Modelo III (com covariáveis: valor médio de IMC)

```
model
{
for (i in 1 : N) {
O[i] ∼ dpois(mu[i])
log(mu[i]) <- log(E[i]) + alpha0 + alpha1 * imc[i] + b[i]
RR[i] <- exp(alpha0 + alpha1 * imc[i] + b[i])
O1[i] ∼ dpois(mu[i])
tp[i]<-O1[i]/PR[i]
}
b[1:N]   car.normal(adj[], weights[], num[], tau)
for(k in 1:sumNumNeigh) {
weights[k] <- 1
}
alpha0   dflat()
alpha1   dnorm(0.0, 1.0E-5)
tau   dgamma(0.5, 0.05)
sigma <- sqrt(1 / tau)
}
```

Análise Bayesiana - Modelo III (dados)

```
list(N = 30,
O = c(1826, NA, NA, 35233, 1842, NA, 3452, NA, 9643, 1951, 1328, 2290, 11906, 1344, 6128, NA, NA, NA, 4221, 1950, 35207,
NA, 2302, 720, 3831, 600, 3885, 2886, 2091, 3792),
E = c( 4280.316824, 4901.15734, 6372.499508, 17838.34636, 6410.199816, 3854.729947, 3437.686045, 4094.987028, 5979.801876,
5619.080017, 4145.312634, 2575.571111, 4899.673362, 909.9561532, 881.759415, 2120.157603, 1530.881776, 1646.723415,
5768.649325, 4037.953121, 30614.5844, 11551.67259, 1801.959501, 2320.303559, 3181.540394, 2302.47843, 4279.516298,
6840.330412, 2696.549893, 2867.133147),
imc=c(-0.9571, 0, 0.5133, -0.1081, -0.0418, 0, -0.4418, 0.3434, 0.8185, 0.3238, 0.3446, 0.3154, -0.3551, 0.9583, 0.2983,
1.49, 0, 0, -3.2398, -0.8959, -0.2459, -0.5954, -1.2043, 0.0761, -1.1295, 0.8068, 0.1806, -0.0814, 1.3278, 1.4993),
PR=c(132830, 192261, 254748, 671340, 254636, 147624, 109093, 119211, 208575, 181582, 140706, 76576, 151787, 23927,
26566, 61387, 42842, 50597, 191401, 125971, 1054768, 409811, 54021, 66149, 93974, 69054, 135438, 230443, 105385, 110168),
num = c(4,4,5,6,6,5,5,5,5,4,5,9,8,5,4,5,5,5,3,3,4,5,7,4,5,5,3,2,0,0),
adj = c(2,8,3,4,
1,3,8,4,
2,4,5,8,1,
2,3,5,6,9,1,
3,4,6,7,8,13,
4,5,7,13,9,
```

6,8,5,13,17,

1,2,3,5,7,

4,6,13,10,12,

9,13,12,11,

10,12,22,23,19,

10,9,13,15,16,14,22,11,18,

10,9,6,5,7,17,15,12,

22,12,16,18,25,

13,17,16,12,

18,14,12,15,17,

7,13,15,16,18,

25,14,17,16,12,

11,23,20,

19,23,21,

20,23,26,24,

25,23,11,12,14,

22,11,19,20,21,26,25,

21,26,27,28,

18,14,22,23,26,

25,23,21,24,27,

26,24,28,

24,27),

sumNumNeigh = 136)

Análise Bayesiana - Modelo III (inits)

list(tau = 1, alpha0 = 0, alpha1 = 0,

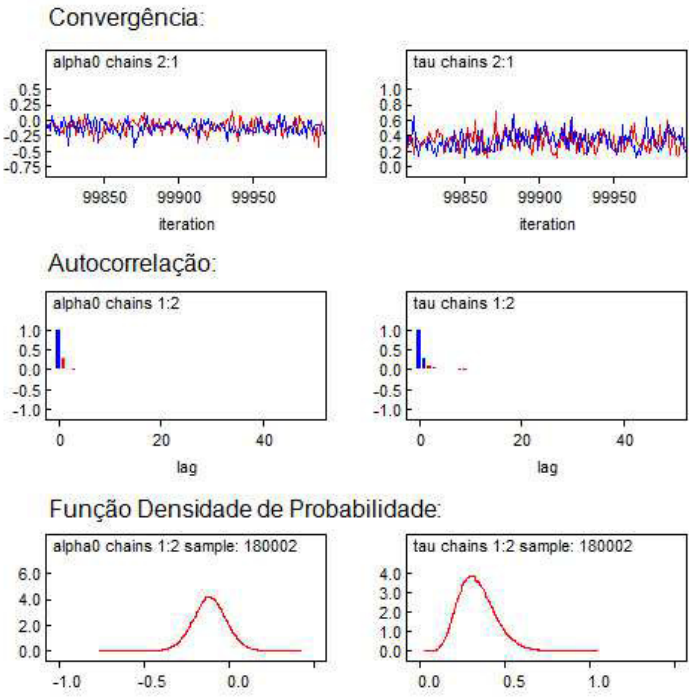
b=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0))

list(tau = 1, alpha0 = 0, alpha1 = 0,

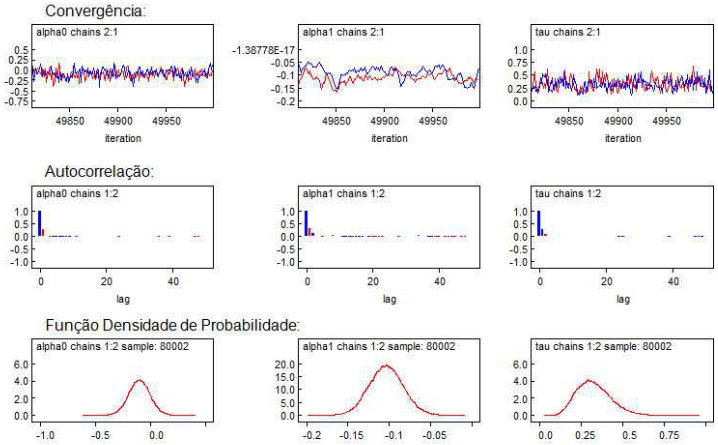
b=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0))

Anexo 3

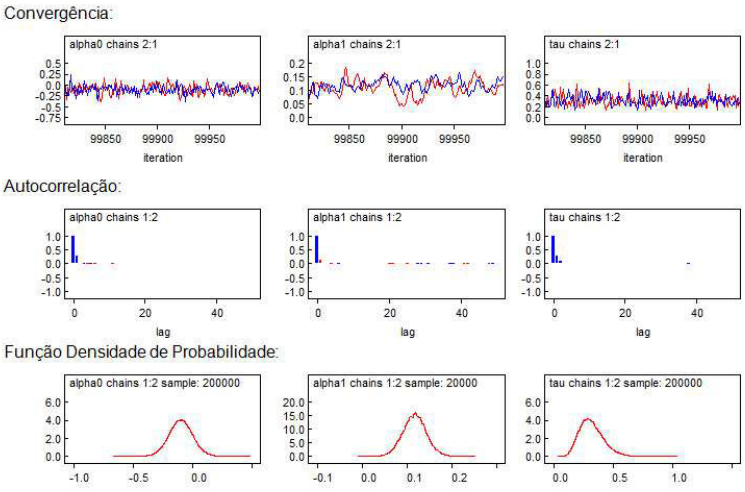
Modelo I - resultados adicionais de convergência dos parâmetros



Modelo II (Idade) - resultados adicionais de convergência dos parâmetros

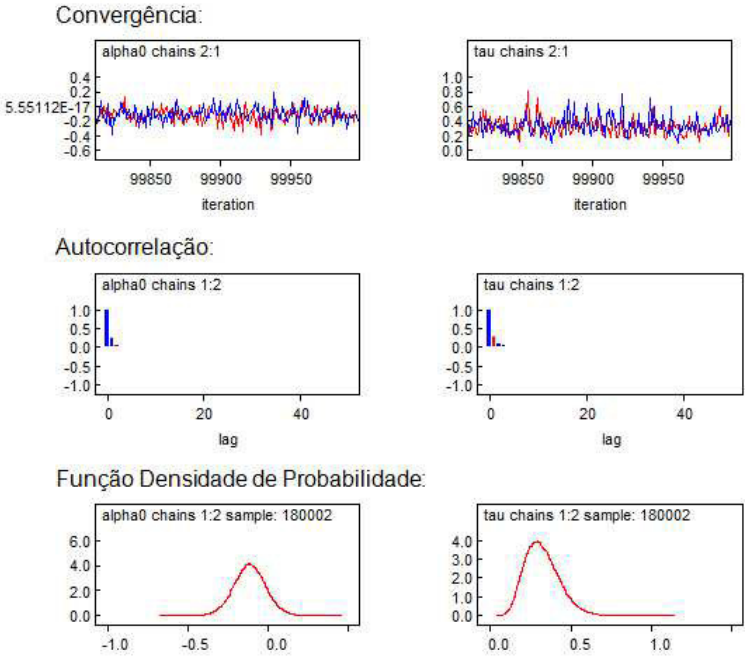


Modelo III (IMC) - resultados adicionais de convergência dos parâmetros

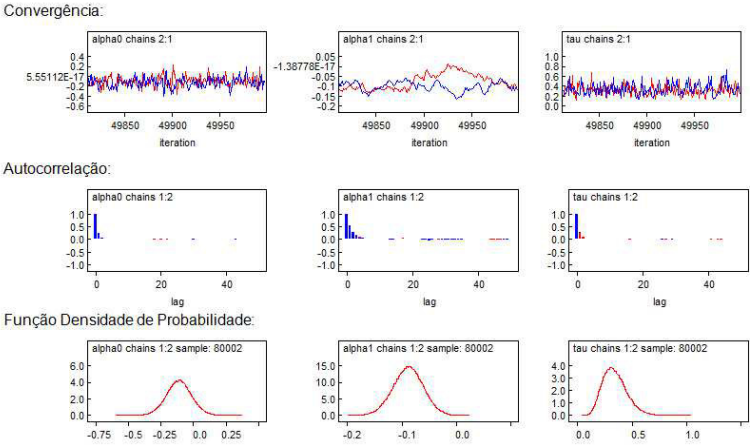


Anexo 4

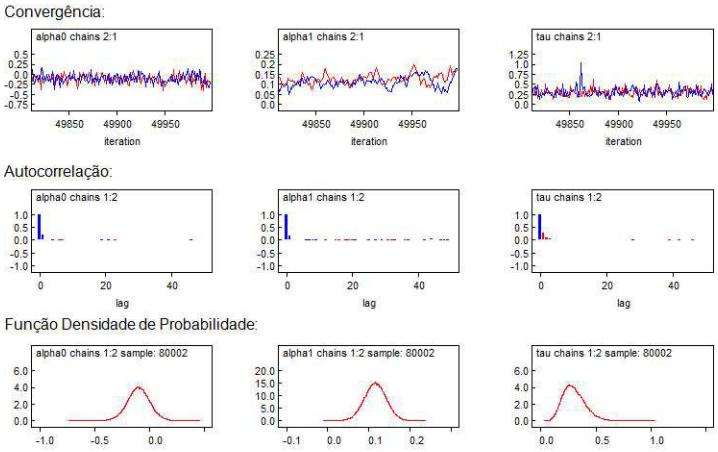
Modelo I - resultados adicionais de convergência dos parâmetros na análise de sensibilidade



Modelo II (Idade) - resultados adicionais de convergência dos parâmetros na análise de sensibilidade



Modelo III (IMC) - resultados adicionais de convergência dos parâmetros na análise de sensibilidade



Bibliografia

Alves, I., Gomes, I. & Sousa, L. Textos de apoio da disciplina de Fundamentos e Metodologias da Estatística. Centro de Estatísticas e Aplicações da Universidade de Lisboa. 2007.

Ascenção, P.(2006). Fibrilhação auricular e prevenção do tromboembolismo - Estudo numa população de Centros de Saúde. Rev Port Clin Geral. 22: 13-24.

Banerjee, S., Carlin, B.P & Gelfan, A.E. (2004). Hierarchical modeling and analysis for spatial data. Chapman & Hall / CRC.

Becker, K. (1999). Bayesian Modelling of Spatial Heterogeneity in Disease Maps with Application to German Cancer Mortality. sonderforschungsbereich. 386: paper 176.

Bell, B.S. & Broemeling, L.D. (2000). A bayesian analysis for spatial processes with application to disease mapping. Stat Med. Apr 15; 19(7): 957-74.

Benjamin E., Levy D., Vaziri S., et al (1994). Independent risk factors for atrial fibrillation in a population-based cohort. JAMA. 271:840-4.

Bernardo, J. M. & Smith, A. F. M. (1994). Bayesian Theory. New York: John Wiley & Sons.

Bernardinelli, L. & Montomoli, C. (1992). Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk. Stat Med. Jun 15; 11(8): 983-1007.

Berger, J. (1985). Statistical Decision Theory and Bayesian Analysis. 2nd ed. Springer-Verlag.

Berger, J. O. & Wolpert, R. (1988). The Likelihood Principle. Second Edition, Hayward, California: Institute of Mathematical Statistics, monograph series.

Besag, J., York, J. & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. Annals of the Institute of Statistical Mathematics. 438(1):1-20.

Biggeri, A., Marchi, M., Lagazio, C., Martuzzi, M. & Böhning, D. (2000). Non-parametric maximum likelihood estimators for disease mapping. Stat Med. Sep 15-30; 19(17-18): 2539-54.

Bonhorst, D., Mendes, M., de Sousa, J., Primo, J., Adragão, P., Andrade, S. & Macedo, A.M. (2010). Epidemiology of atrial fibrillation. Rev Port Cardiol. 29(7-8): 1207-17.

Bonhorst, D., Mendes, M., Adragão P., De Sousa, J., Primo, J., Leiria, E. & Rocha, P. (2010). Prevalence of atrial fibrillation in the Portuguese population aged 40 and over: the FAMA study. Rev Port Cardiol. 29(3): 331-50.

Borgoni, R. & Billari, F. (2003). Bayesian spatial analysis of demographic survey data: An application to contraceptive use at first sexual intercourse. Demographic Research. Vol.8, pp.61-92.

Breslow, N.E. & Day, N.E. (1975). Indirect standardization and multiplicative models for rates, with reference to the age adjustment of cancer incidence and relative frequency data. J Chronic Dis. Jun; 28(5-6): 289-303.

Brook, S.P. & Gelman, A. (1997). General Methods for Monitoring Convergence of Iterative Simulations. Journal of Computational and Graphical Statistics. 7, 434-455.

Carlin, B. & Louis, T. (2000). Bayes and Empirical Bayes Methods for Data Analysis, 2nd ed. Chapman & Hall / CRC.

Carvalho, M.L. Textos de apoio da disciplina de "Introdução à Epidemiologia". Faculdade de Ciências da Universidade de Lisboa. 2007.

Carvalho, M. L. & Natário, I.C.(2008). Análise de Dados Espaciais. Editores: Sociedade Portuguesa de Estatística.

Chien, K., Su, T., Hsu, H., et al (2008). Atrial fibrillation prevalence, incidence and risk of stroke and all-cause death among Chinese. Int J Cardiol. Nov; 29.

Clayton, D. & Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics. 43(3): 671-81.

Costa, A.C. (2000). Técnicas de estimação no âmbito da pós-estratificação. Dissertação para tese de mestrado não publicada. Instituto Superior de Estatística e Gestão de Informação, Universidade Nova de Lisboa.

Cowles, M. K. & Carlin, B. P. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. Journal of the American Statistical Association, 883-904.

Dellaportas, P., Forster, J.J. & Ntzoufras, I. (2002). Joint Specification of Model Space and Parameter Space Prior Distributions. Statist. Sci. Volume 27, Number 2 (2012), 161-318.

Feinberg, W.M., Blackshear, J.L., Laupacis, A., Kronmal, R. & Hart, R.G. (1995). Prevalence, age distribution, and gender of patients with atrial fibrillation. Analysis and implications. Arch Intern Med. Mar 13;155(5):469-73.

Ferreira, P.S. (2008). Fibrilhação Auricular: tromboprofilaxia e controvérsia na prática clínica. Dissertação para tese de mestrado não publicada. Faculdade de Ciências da Saúde, Universidade da Beira Interior.

García-Acuña, J.M., González-Juanatey, J.R., Ezquerra, E.A., et al (2002). Permanent atrial fibrillation in heart disease in Spain. The CARDIOTENS Study 1999. Rev Esp Cardiol 2002.

55(9):943-52.

Gelfand, A. & Vounatsou, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. Biostatistics. 4:11-15.

Gelman, A., & Rubin, D.B. (1992). Inference from Iterative Simulation Using Multiple Sequences. Statistical Science. 7, 457-511.

Geweke, J. (1992). "Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments."In Bayesian Statistics 4 (eds). J.M. Bernardo, J. Berger, A.P., Dawid & A.F.M. Smith. Oxford: Oxford University Press, 169-193.

Gordis, L. (1996). Epidemiology. Elsevier - Health Sciences Division.

Gomes-Rubío, V. & López-Quílez, A. (2000). Empirical and Full Bayes estimators for disease mapping. Acedido em www.unavarra.es/metma3/Papers/PDFS.../gomezrubiovirgilio2.pdf, a 2012/11/30.

Greenland, S. & Rothman, K.J. (2008). Measures of Occurrence. In Rothman, K.J., Greenland, S. & Lash, T.L., Modern Epidemiology. Third Edition (pp. 32-50). Lippincott Williams & Wilkins.

Heidelberger, P. & Welch, P.D. (1983). Simulation Run Length Control in the Presence of an Initial Transient. Operations Research. 31: 1109-1144.

Hegarty, A.C., Carsin, A.E. & Comber. H. (2010). Geographical analysis of cancer incidence in Ireland: a comparison of two Bayesian spatial models. Cancer Epidemiol. 2010 Aug;34(4):373-81

Kafadar, K. (1997). Geographic trends in prostate cancer mortality: an aplication of spacial smoothers and the need for adjustment. Ann Epidemiol. 7; 35-45.

Kaldor, J. & Clayton, D. (1989). Role of advanced statistical techniques in cancer mapping. Recent Results Cancer Res. 114:87-98.

Kannel, W.B. & Benjamim, E.J. (2008). Final draft status of the Epidemiology of Atrial Fibrilation. Med Clin North Am. January; 92(1): 17-ix.

Knorr-Held, L. & Best, N. G. (2001), A shared component model for detecting joint and selective clustering of two diseases. Journal of the Royal Statistical Society: Series A (Statistics in Society), 164: 73-85.

Lawson, A. (2009). Bayesian disease mapping: hierarchical modeling in spatial epidemiology. Chapman & Hall / CRC.

Lawson, A., Browne, W. & Rodeiro, C. (2003). Disease Mapping with WinBugs and MlwiN. John Wiley & Sons, Ltd.

Lawson, A., Biggeri, A., Böhning, D., Lesaffre, E., Clark, A., Schlattmann, P. & Divino, F. (2000). Disease mapping models: an empirical evaluation. Statistics in Medicine 2000; 19: 2217-41.

Lima-Costa, M.F. & Barreto, S.M (2003). Tipos de estudos epidemiológicos: conceitos básicos e aplicações na área do envelhecimento. Epidemiologia e Serviços de Saúde 2003; 12(4):189 - 201.

Lunn, D., Spiegelhalter, D., Thomas, A. & Best, N. (2009). The Bugs Project: evolution, critique and future directions. Statistics in Medicine. 28:3049-3067.

Marshall, R.J. (1991). Mapping disease and mortality rates using empirical Bayes estimators. J R Stat Soc Ser C Appl Stat. 40(2): 283-94.

MacNab, Y.C., Farrell, P.J., Gustafson, P. & Wen, S. (2004). Estimation in Bayesian disease mapping. Biometrics. Dec; 60(4): 865-73.

Miyasaka, Y., Barnes, M.E., Gersh, B.J., et al. (2006). Secular trends in incidence of atrial fibrillation in Olmsted County, Minnesota, 1980 to 2000, and implications on the projections for future prevalence. Circulation. Jul 11;114(2):119-25.

Natário, I. (2004). Hierarchical Bayesian Models for Epidemiological Analysis for Rare Events. Dissertação de tese de Doutoramento não publicada, Faculdade de Ciências da Universidade

de Lisboa, Lisboa, Portugal.

Prieto, R.R., García-Pérez, J., Pollán, M., et. al (2007). Modelling of municipal mortality due to haematological neoplasias in Spain. J Epidemiol Community Health 2007;61:165-171.

Raftery, A.E. & Lewis, S. (1992). How Many Iterations in the Gibbs Sampler? In Bayesian Statistics 4 (eds. J.M. Bernardo, J. Berger, A.P. Dawid and A.F.M. Smith). Oxford: Oxford University Press, 763-773.

Robert, C.P. (2001). The Bayesian Choice, Second Edition, New York: Springer-Verlag.

Richardson, S., Thomson, A., Best, N. & Elliott, P. (2004). Interpreting posterior relative risk estimates in disease-mapping studies. Environ Health Perspect. Jun;112(9):1016-25.

Schokkaert, E. % Van de Voorde, C. (2009). Direct versus indirect standardization in risk adjustment. J Health Econ. Mar;28(2):361-74.

Spiegelhalter, D.J., Thomas, A., Best, N. & Lunn, D (2002). WinBUGS User Manual, Version 1.4.

Strong, M., Pearson, T., MacNab, Y.C. & Maheswaran, R. (2012). Mapping gender variation in the spatial pattern of alcohol-related mortality: a Bayesian analysis using data from South Yorkshire, United Kingdom. Spat Spatiotemporal Epidemiol. Jun; 3(2): 141-9.

Torabi, M. (2013). Hierarchical Bayesian bivariate disease mapping: analysis of children and adults asthma visits to hospital. Journal of Applied Statistics. Vol. 41, No. 3, 612-621.

Tsutakawa, R.K. (1985). Estimation of cancer mortality rates: a Bayesian analysis of small frequencies. Biometrics. 41(1):69-79.

Tsutakawa, R.K., Shoop, G.L. & Marienfeld, C.J. (1985). Empirical Bayes estimation of cancer mortality rates. Stat Med. 4(2):201-12.

Wasserman, L. (2004). All of Statistics: A Concise Course in Statistical Inference. New York: Springer-Verlag.

Whilhelmsen, L., Rosengren, A. & Lappas, G. (2001). Hospitalizations for atrial fibrilation in the general male population: morbidity and risk factors. Journal of Internal Medicine. 250:382-389.

Wolf P., Abbott R. & Kannel W. (1991). Atrial fibrillation as an independent risk factor for stroke: The Framingham Study. Stroke. 22:983-8.